Cognitive Psychology Unit

Chunking and Rehearsal in Working Memory: A Matter of Central Attention?

Thesis (Cumulative Thesis)

Presented to the Faculty of Arts and Social Sciences

of the University of Zurich

for the degree of Doctor of Philosophy (PhD)

by Mirko Thalmann

Accepted in the fall semester 2017

on the recommendation of the doctoral committee:

Prof. Dr. Klaus Oberauer (Main Supervisor)

Prof. Dr. Edward Awh

Zurich, 2017

Abstract

Several theories of working memory state that central attention has an important role in the maintenance of representations over the short term. Here, I examine two propositions how central attention is used in working-memory tasks. The first proposition states that central attention allows people to focus on four chunks of information simultaneously while not focused information is forgotten. The ability to remember information over the short term should be independent of the size of chunks. That was not confirmed in Study 1, rendering the idea unlikely that working memory is limited by a fixed number of chunks. The second proposition states that two reactivation mechanisms – refreshing and articulatory rehearsal – can be carried out in parallel to counteract memory decay because only refreshing requires central attention. Study 2 did not confirm this prediction. It showed that articulatory rehearsal also requires central attention. Therefore, it is implausible that refreshing and articulatory rehearsal can be carried out simultaneously without any cost. The two studies question whether central attention is important in the maintenance of information over the short term.

Available statistics packages were not able to analyze the data of Studies 1 and 2 with state-of-the-art methods. Therefore, I have developed a new statistics package called BayesRS. In Study 3, I tested the functionality of the new package. The results showed that BayesRS represents a viable tool to analyze experimental data. By accounting for the structure in the data adequately, BayesRS prevents premature conclusions about the existence of effects.

Zusammenfassung

Verschiedene Theorien des Arbeitsgedächtnisses gehen davon aus, dass zentrale Aufmerksamkeit wichtig fürs kurzzeitige Merken von Informationen ist. Ich untersuche in dieser Arbeit zwei theoretische Ideen, wie zentrale Aufmerksamkeit fürs kurzzeitige Merken eingesetzt wird. Die erste Idee geht davon aus, dass man die Aufmerksamkeit gleichzeitig auf vier Chunks richten kann. So behält man diese vier Chunks im Arbeitsgedächtnis, währenddem nicht fokussierte Informationen vergessen gehen. Die Merkfähigkeit sollte dabei nicht von der Grösse der Chunks abhängen. Studie 1 konnte diese Idee aber nicht bestätigen, weshalb es unwahrscheinlich ist, dass das Arbeitsgedächtnis durch eine Anzahl Chunks begrenzt wird. Die zweite Idee geht davon aus, dass zwei Reaktivierungsmechanismen – artikulatorisches Rehearsal und Refreshing – gleichzeitig ausgeführt werden können, um dem Zerfall von Gedächtnisspuren entgegenzuwirken. Das parallele Ausführen der Mechanismen ist möglich, weil davon ausgegangen wird, dass nur Refreshing, aber nicht artikulatorisches Rehearsal zentrale Aufmerksamkeit benötigt. Studie 2 konnte diese Idee nicht bekräftigen. Die Ergebnisse zeigen, dass artikulatorisches Rehearsal entgegen der Annahme trotzdem zentrale Aufmerksamkeit benötigt. Artikulatorisches Rehearsal und Refreshing können deshalb nicht gleichzeitig ausgeführt werden ohne Kosten. Die beiden Studien werfen die Frage auf, ob zentrale Aufmerksamkeit tatsächlich wichtig fürs kurzzeitige Merken ist.

Die Daten von Studien 1 und 2 konnten mit verfügbaren Statistikpaketen nicht auf dem neusten Stand der Statistik analysiert werden. Deshalb habe ich ein neues Statistikpaket mit Namen BayesRS entwickelt. In Studie 3 habe ich die Funktionsfähigkeit von BayesRS getestet. Die Ergebnisse zeigen, dass BayesRS gut geeignet ist, experimentelle Daten zu analysieren. Indem es möglich ist, die Struktur der Daten adäquat im statistischen Modell abzubilden, verhindert es vorschnelle Schlüsse über das Vorhandensein von Effekten.

Contents

PART I: SYNOPSIS

1. Introduction

Baddeley and Hitch (1974) were the first to show that performance in reasoning and text comprehension tasks decreases when people have to concurrently remember a few digits in their serial order. Because the same system was supposed to be involved in reasoning/comprehension and in retention of the digits they called it working memory (WM), in contrast to the previously used term short-term memory. Since then, the number of investigations about WM has increased dramatically. Not least because WM has proofed to be amongst the best predictors of variables such as mathematical achievement (Bull, Espy, & Wiebe, 2008), academic achievement in general (Alloway & Alloway, 2010), reading comprehension (Daneman & Carpenter, 1983), and language learning (Baddeley, Gathercole, & Papagno, 1998). WM also correlates highly with fluid intelligence (Süß, Oberauer, Wittmann, Wilhelm, & Schulze, 2002), which led some authors to conclude that fluid intelligence is just little more than WM capacity (Kyllonen & Christal, 1990).

One of the central characteristics of WM is that its capacity is severely limited. For example, when people are required to remember a random configuration of visually presented objects, they can only remember about four of them in their correct location (Luck & Vogel, 1997). A whole category of WM theories assumes that the observation of the capacity limit is due to a limited resource (for an overview of all categories see Oberauer, Farrell, Jarrold, & Lewandowsky, 2016). Often, this resource is supposed to be central attention. However, how central attention is actually used in WM and how it may explain the capacity limit differs largely between theories. In the first two studies of this thesis I examine two theoretical conceptions how central attention is used in WM.

Cowan (1999, 2001) assumes that central attention, or in his words the focus of attention, allows people to keep four chunks active in WM. In that case, the resource – central attention – is subdivided into four equal parts and each part can remember one chunk. In Study 1, I tested whether WM capacity can be well described by a fixed number of chunks. A

different conception how central attention is used in WM is put forward by Barrouillet and colleagues in the time-based resource-sharing model (TBRS, Barrouillet, Bernardin, & Camos, 2004; Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007). These authors assume that central attention counteracts the inexorable decay of memory traces by a process called attentional refreshing. Central attention is characterized as a structural bottleneck that enforces serial processing (Pashler, 1994). Therefore, only one representation can be refreshed at once and refreshing has to cycle around constantly within a set of representations. Recently, articulatory rehearsal was added as a second restoration mechanism to the TBRS model (Camos, Lagner, & Barrouillet, 2009). Crucially, it is assumed that articulatory rehearsal can be carried out in parallel with refreshing without any costs. This hinges on the assumption that articulatory rehearsal does not require central attention. I tested this assumption in Study 2.

The third study of the thesis presents a new statistics package called BayesRS and tests the functionality of the new package. I have developed BayesRS as a tool to analyze the data of the first two studies in the thesis. The package accepts normally distributed data as the dependent variable and computes default Bayes factors (BFs) for categorical and continuous predictor variables in hierarchical Bayesian models. As a new feature, it allows users to compute BFs for continuous predictors including individual differences in their effect size (i.e., random slopes), which has not been possible in previous Bayesian statistics packages. Thereby, it brings together two fields that have been argued to be of great advantage for psychological research: mixed-effects modeling and Bayesian statistics.

In the first part of the thesis, I am summarizing and discussing the results of the three studies. For the better understanding, I am going to start each section with an introduction to the relevant theoretical background. In the second part, the manuscripts of the three studies are printed in full length.

2. Study 1: “How Does Chunking Help Working Memory?”

2.1. Cowan’s embedded-processes model

Cowan (1999) assumes that WM is organized as hierarchically arranged faculties (see Figure 1). Information that is presented and associations to this information are represented in the activated portion of long-term memory (LTM). They are assumed to be in a heightened state of activation, but they are vulnerable to the mechanisms of decay and interference. At the same time, four chunks can be brought to the focus of attention, which is capacity-limited. Representations in the focus of attention are available to conscious awareness (Cowan, 2001) and immediately available for processing (Gilchrist & Cowan, 2011). With other words, the capacity limit is due to the limited resource central attention, which is assumed to be discretized into four slots and each slot can hold active one chunk. Processing strategies (e.g., mnemonic strategies) can increase the amount of information that can be recalled in a given task over the capacity of the focus of attention. Critically, these strategies have to be disabled to assess the pure capacity limit of the focus of attention.

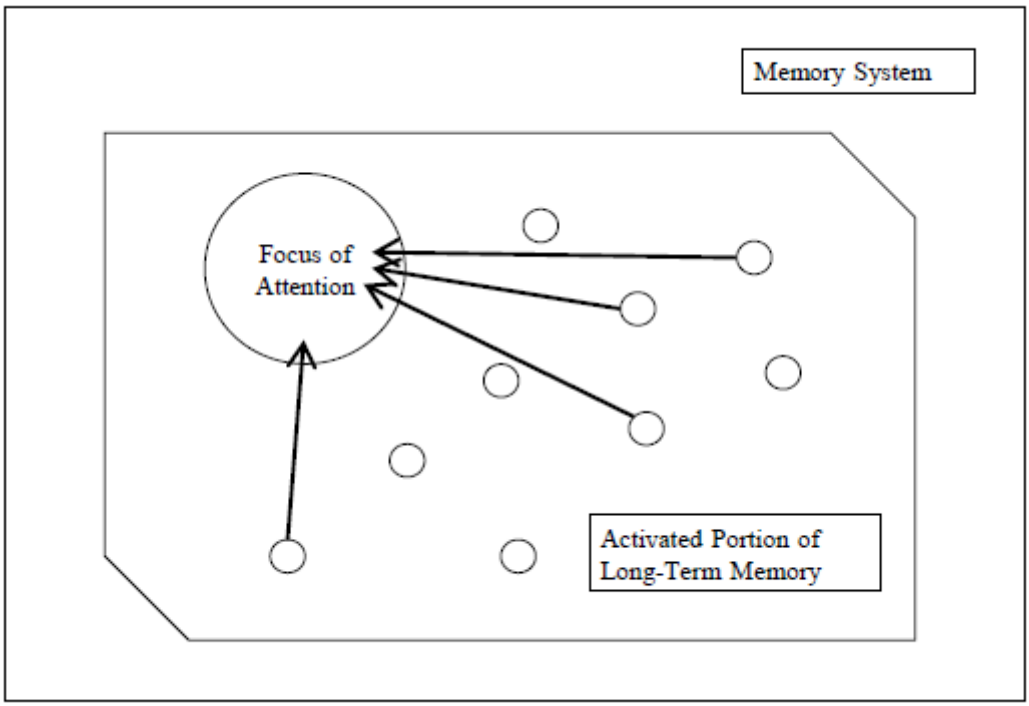

Figure 1. Hierarchy of the WM system in Cowan's embedded-processes model. Adopted from Figure 1 in Cowan (2001).

The concept of chunks is given a special role in the embedded-processes model. By definition of Cowan (2001), a chunk consists of a set of representations that have strong associations to each other but no or only weak associations to other representations. The associations within a chunk are assumed to be stored in LTM. An often used example for a chunk is the acronym FBI, consisting of a well learned series of three letters, but a chunk can also be a word, consisting of a well learned series of several syllables. By stating that the focus of attention is limited by the number of chunks, Cowan assumes that the size of chunks is irrelevant (Cowan, 2001).

Previous studies examining chunking in the context of the embedded-processes model (Chen & Cowan, 2005, 2009; Cowan, Chen, & Rouder, 2004) supported the prediction that people can remember a fixed number of chunks in WM tasks. For example, in Chen and Cowan (2005) participants were required to remember a series of words either consisting of learned word pairs (i.e., chunks) or of random word pairs for immediate recall. The

participants remembered about twice as many words when the series consisted of word pairs as when it consisted of words. Therefore, Chen and Cowan argued that WM capacity is well described by the number of chunks, but not by the size of the chunks. Although these results are consistent with the idea that the focus of attention operates on chunks (i.e., chunks are the basic storage unit in WM), they can also be explained by a different mechanism. It could be that learned word pairs are remembered in the same way as random word pairs. But then, at the time of recall information from LTM can assist in the reconstruction of the learned pairs but not in the reconstruction of random pairs (Hulme, Maughan, & Brown, 1991; Hulme, Roodenrys, Brown, & Mercer, 1995). A special experimental set-up common in all three chunking studies further complicates their interpretation. That is, chunks consisted of unique elements that could not be part of other lists. It was therefore immediately clear for the participants after presentation of the first element whether a chunk would follow and which chunk it would be. As a consequence, participants could strategically restrict encoding and maintenance to the first element of the chunk. It is therefore unclear whether the results from these previous chunking studies generalize to circumstances when this strategy cannot be applied.

2.2. Summary of Study 1

In Study 1, I examined two predictions from the embedded-processes model, and I investigated an additional question about chunking. First, I tested whether chunking frees capacity. For that purpose, I focused on the recall of not-chunked information in all experiments to exclude the possibility that the beneficial effect of chunks arises due to LTM assistance at the time of recall. For example, I tested whether items 4-6 in the two lists F B I S H K (list 1) and I F B S H K (list 2) are remembered better in the context of a chunk (F B I in list 1) than in the context of a random sequence (I F B in list 2). Second, I tested whether the potential chunking benefit is independent of the size of the chunk. For example, I tested

whether items 5-7 in the list H T M L S H K are remembered equally well as items 4-6 in the list F B I S H K. Finally, I investigated how the chunking benefit is affected by the serial position of the chunk within a list to be remembered.

In agreement with the embedded-processes model I observed that chunks free capacity in WM. This was the case in two experimental paradigms using different stimulus materials. I observed, however, a difference in the size of the chunking benefit across the two paradigms. When chunks consisted of unique stimuli the chunking benefit was independent of chunk size confirming the prediction from the embedded-processes model. However, when chunks were constructed with overlapping stimuli that could also appear in other lists the benefit was smaller for large than for small chunks, which is against the prediction of the embedded-processes model. The model might explain this effect by assuming that people (sub-) vocally rehearsed all the individual items presented in a list. Because rehearsal of large chunks takes longer than rehearsal of small chunks, not-chunked items would be rehearsed less often in the context of large chunks than in the context of small chunks. I excluded this explanation in one of the experiments, in which participants were instructed to repeatedly articulate unrelated syllables aloud (i.e., “ba bi bu…”, articulatory suppression, AS) during presentation and retention. This measure is supposed to block articulatory rehearsal of the memoranda (Richardson & Baddeley, 1975).

I suggest that the difference between small and large chunks is due to the following two reasons. First, with unique stimuli, participants only have to encode and remember the first item of a chunk. With overlapping stimuli, this strategy does not work. Instead, people may represent the chunk in WM differently, for example with a semantic representation retrieved from LTM (Ericsson & Kintsch, 1995), which may sometimes not be available at all or not be available quickly enough. Second, larger chunks could be represented by more complex representations, which would lead to more interference with the representations of the not-chunked items in a trial (Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012).

Finally, I observed that the chunking benefit depended on the serial position of the chunk within a list. When a chunk was presented in the beginning of a list, all later presented items benefitted from the chunk. The same was true for a chunk in the middle of a list. In addition, the midway-presented chunk improved memory for items presented before. A chunk presented last in a trial did not improve memory for the previously presented items. I assume that the mechanism leading to the chunking benefit includes three steps. First, the individual items of the chunk are encoded into WM. Second, a match between the currently active representations in WM and the chunk representation in LTM can lead to retrieval of the chunk representation into WM. Third, the representations of the individual items have to be removed from WM. Especially the last step may depend on the load that is already imposed on WM. That is, removal is easiest in the beginning of a trial, because it can be achieved in one rapid sweep (Ecker, Oberauer, & Lewandowsky, 2014). Later on in the trial, representations of the individual chunk items interfere with representations of previously presented items (Oberauer et al., 2012) and they have to be removed individually, which is arguably more difficult and error-prone than a sweep.

2.3. Implications for the embedded-processes model

The consistent finding of a chunking benefit in four experiments confirms that the concept of a chunk is useful in explaining WM capacity limitations. However, the fact that the chunking benefit varied as a function of chunk size cannot be accommodated with the embedded-processes model in a straightforward way. That is, the capacity of the focus of attention is assumed to be limited by four chunks, independent of their size. Hence, the current results have to be explained by referring to something else than the focus of attention. Two potential candidates outside the focus of attention are rehearsal counteracting decay and interference (Cowan, 2001).

As argued above, it is unlikely that additional rehearsal can explain the pattern of results. That is because the chunking benefit depended on chunk size even when participants had to do AS. In contrast, interference between representations in WM in combination with removal of representations from WM can explain the results of Study 1. Hence, the embedded-processes model could accommodate the current results when relying on interference and removal happening outside the focus of attention. This explanation is, however, not straightforward because removal is likely to require selection of representations in WM, which is arguably accomplished by the focus of attention (Oberauer, 2009). A further drawback of this explanation is that the focus of attention itself does not have explanatory power anymore when capacity limitations are explained by referring to mechanisms outside the focus of attention. It can be doubted, therefore, whether the conceptualization of WM as the limited resource central attention is actually theoretically useful when the results can be explained by solely relying on different mechanisms (Navon, 1984).

2.4. Implications for theories of working memory in general

The results of Study 1 align nicely with studies showing that LTM influences performance in WM tasks. For example, Hulme et al. (1991, 1995) showed that a higher memory span for words than for non-words is likely due to a LTM contribution to WM. They suggested that LTM assists in the reconstruction of partially degraded representations at the time of recall. Study 1 extends these results because it shows that LTM influences WM performance not only during recall but also during maintenance. Together, the results favor models that describe WM in close relation to LTM. For example, the current results can be explained straightforwardly within the three-embedded components model by Oberauer (2002, 2009). This model assumes a similar memory hierarchy as Cowan (1999, 2001). But in contrast to Cowan's model, it does not refer to a limited resource to explain the capacity limitation. Instead, it explains the observation of a capacity limit with mechanisms of

interference (retrieval competition and representational interference). The focus of attention selects representations in WM for further processing, which is for example required in the updating of old WM contents with new WM contents. Therefore, the model provides a straightforward explanation for the current results.

A different model that can explain how chunks, which are part of a list, can be recognized is the model by Page and Norris (2009). One appealing feature of that model is that it can learn chunks consisting of different numbers of items. This allows the model to recognize chunks that do not comprise a whole list themselves. It is unlikely, however, that the Page and Norris model is able to predict the pattern of the chunking benefit across serial positions as observed in Study 1. The reason why that is unlikely is that the model incorporates rehearsal as a mechanism to counteract decay. For example, rehearsing a three-item chunk requires recalling all constituents of the chunk. Decay of the representations of the not-chunked items is therefore independent of whether three items constitute a chunk or not. In contrast, decay theories assuming that a three-item chunk can be rehearsed more quickly than three individual items, predict that chunks should help in any serial position. That is because not-chunked items are expected to decay less during retention when a chunk can be rehearsed more quickly. This prediction was clearly not supported by the current data.

2.5. Implications for the understanding of central attention

What tells us Study 1 about the use of central attention in WM? I tested the assumption of the embedded-processes model (Cowan, 1999, 2001) that central attention is used to remember four chunks in WM, independent of their size. Although I consistently observed a chunking benefit, it depended on the size and on the serial position of the chunk. This renders it implausible that central attention keeps active a fixed number of chunks and we may think of different ways central attention is used in WM.

One different conceptualization is inspired by Garavan (1998) who let participants update counters of geometric figures that were sequentially presented on the screen. The results showed that repeatedly updating the same counter takes less time than updating a counter that was not updated previously. This led Garavan to conclude that attention is used to focus on one single representation at a time in WM. This contrasts with Cowan's idea that several chunks/representations can be accessed simultaneously. Accessing a single representation in WM can also be considered to be advantageous because it prevents confusion between several representations (Oberauer & Bialkova, 2009). Arguably, many tasks require access to individual representations (e.g., mental arithmetic).

3. Study 2: "Revisiting the Attentional Costs of Rehearsal in Working-Memory Tasks"

3.1. The time-based resource-sharing model

The time-based resource-sharing model of WM (TBRS, Barrouillet et al., 2004; Barrouillet, Bernardin, Portrat, Vergauwe, & Camos, 2007) is mainly inspired by the complex-span paradigm. In this task, the presentation of memoranda is interleaved with sequences of a processing task, for example requiring participants to judge the accuracy of simple equations (Turner & Engle, 1989). The TBRS model assumes that the activation of representations in WM decays over time. Decay can be counteracted by refreshing partially degraded representations. Crucially, refreshing and processing are assumed to require central attention. Because central attention is considered to rely on a structural response-selection bottleneck (Pashler, 1994) only refreshing or processing can be carried out at once. This leads to a constant switching between processing and refreshing when a series of processing tasks has to be carried out (see Figure 2). In support of their theory, Barrouillet et al. (2007) showed that memory performance in the complex-span paradigm is negatively related to the proportion of time required for responding in the processing task compared to the total available processing time (i.e., the cognitive load, CL). That is, memory gets worse the more of the available time is used for processing.

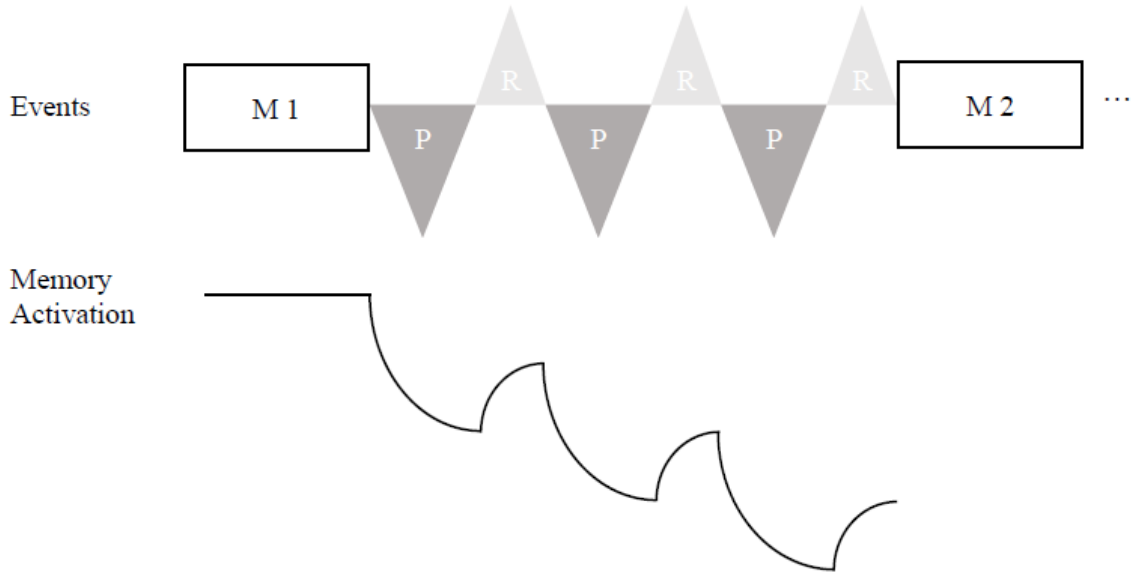

Figure 2. Interplay between decay and refreshing in the TBRS model. M1 and M2 represent the presentation of two memory items. Memory activation of M1 decays during processing (P) but can be refreshed (R) between two processing episodes. Figure adopted from Figure 1 in Oberauer and Lewandowsky (2014).

Recently, Camos, Lagner, & Barrouillet (2009) added articulatory rehearsal as a second restoration mechanism to the TBRS model. Importantly, they concluded that articulatory rehearsal and refreshing are independent mechanisms leading to additive effects on memory performance (see also Camos & Barrouillet, 2014). Whereas refreshing is attentionally demanding per definition, rehearsal is not assumed to require central attention at all. The TBRS model predicts that it is possible to carry out the two restoration mechanisms simultaneously without any cost. Vergauwe, Camos, and Barrouillet (2014) examined this prediction by presenting participants with a varying number of letters to remember and requiring them to respond to a series of processing tasks during retention. They observed that reaction times (RTs) in the processing task increased linearly with memory set size when people were required to do AS during retention. In contrast, without AS the increase of processing RTs with memory set size was negligible. Vergauwe and colleagues interpreted these results in such a way that participants use articulatory rehearsal as a default rehearsal mechanism. People only switch to refreshing when articulatory rehearsal is not possible, that

is, when they are required to do AS during retention. This would explain the substantial increase of processing RTs with memory set size in the AS condition and the absence of such an increase in the condition without AS. What remains unclear in the study by Vergauwe and colleagues, however, is whether participants actually used articulatory rehearsal to remember the letters in the no AS condition. As there was no means to ascertain that, the study cannot conclude that articulatory rehearsal can be carried out without any central attentional costs.

3.2. Summary of Study 2

In Study 2, I revised whether articulatory rehearsal can be carried out without any central attentional costs as predicted by the TBRS model. For that purpose, I used similar tasks as Vergauwe and colleagues. Participants were presented with a variable number of words to be rehearsed aloud (Rundus & Atkinson, 1970; Tan & Ward, 2008). This overt rehearsal method allowed to control that participants actually rehearsed the words. While rehearsing, participants had to respond to a series of processing tasks requiring central attention. Analysis of RTs in the processing task in relation to the number of rehearsed words is informative about whether rehearsal demands central attention.

Carrying out articulatory rehearsal delayed concurrent processing for at least 10 s. The average RT cost of each additionally to be rehearsed item is approximately 20-25 ms. Analyses with a reaction-time model allowed us to unambiguously attribute the slowing in the processing task to central attentional demands of articulatory rehearsal. The model-based analyses also revealed that the central attentional demand of articulatory rehearsal increases disproportionately at set sizes 3 and 4. The attentional demand of rehearsing only one or two words must be fairly small. A potential alternative explanation of these results is that participants, in anticipation of a WM test, use an additional rehearsal mechanism to remember the words when instructed to rehearse aloud. The measured costs would then not be due to

articulatory rehearsal, but due to the additionally used rehearsal mechanism. In an additional experiment, I excluded this possibility.

In this experiment, I examined whether people additionally use refreshing or elaboration when instructed to rehearse aloud. Elaboration is defined as the enriching of to be remembered representations with representations from LTM (Craik & Tulving, 1975), which could also be the integration of mental images of to be remembered items (Bower, 1970). First, the results render it unlikely that people additionally refresh when they have to rehearse a few words. To come to this conclusion, I compared the rehearse-aloud condition to a condition, in which participants were instructed to carry out AS during retention. In contrast to the rehearsal condition, there were no persistent costs on processing RTs in the AS condition. If participants were refreshing in the rehearsal condition, they should refresh in the AS condition as well, because in this condition the use of rehearsal is blocked. In addition, memory in a surprise delayed recognition test was better in the AS condition than in the rehearsal condition. If anything, the reversed pattern should have resulted if people only refreshed in the rehearsal condition because refreshing but not rehearsal is assumed to improve long-term memory (Camos & Portrat, 2015). Second, it is neither likely that people elaborate in addition to articulatory rehearsal because delayed recognition was much worse in the articulatory rehearsal condition than in an instructed elaboration condition, in which participants were instructed to generate interactive images of the presented words. Furthermore, the pattern of the costs on processing RTs differed substantially between the instructed rehearsal and the instructed elaboration condition. Processing RTs decreased heavily over the processing period in the elaboration condition but remained constant in the rehearsal condition.

3.3. Implications for the time-based resource-sharing model

The TBRS model predicted that refreshing and articulatory rehearsal can be carried out simultaneously. Study 2 showed that articulatory rehearsal has central attentional costs. Therefore, the results rule out that refreshing and articulatory rehearsal can be carried out simultaneously without any additional attentional costs. For example, let us assume that three letters can be rehearsed within about one second. That means that central attention is required for approximately 70 ms to do so. Only the remaining 930 ms (i.e., 93 %) can be used for refreshing. Admittedly, this does not represent a big loss due to the attentional demands of rehearsal. This may, however, not be the only central attentional cost when people are required to coordinate refreshing and articulatory rehearsal. Carrying out articulatory rehearsal and refreshing simultaneously represents a dual-task situation. It has been shown that the requirement to coordinate two tasks itself is attentionally demanding (Logie, Cocchini, Della Sala, & Baddeley, 2004). Therefore, the coordination costs have to be added to the costs of rehearsal, and they will further slow refreshing down.

An additional finding of Study 1 complicates the interpretation of the results within the TBRS model: We did not observe persistent central attentional costs in the AS condition, in which participants are supposed to use refreshing according to the TBRS model. The representations of the memoranda should decay and be forgotten because rehearsal was blocked by AS and people did not continuously refresh the words. Therefore, in the context of the TBRS model it remains unclear why people were able to remember up to four words in their correct serial order.

The conclusion by Vergauwe et al. (2014) that refreshing only intervenes when articulatory rehearsal is not possible may be a larger revision of the TBRS model than it appears at first sight. It predicts that there will be no CL effect in complex-span tasks with a rather small verbal memory load and a non-verbal processing task. Experiments 2 and 5 in Barrouillet, Plancher, Guida, and Camos (2013) come closest to a test of this prediction: In a

complex-span task, participants were required to remember five letters and to give an answer on the keyboard in a visual-spatial processing task. In both experiments, a CL effect was observed. That is, memory for the five letters decreased with increasing pace of the processing task. These results do not confirm the prediction from the TBRS model. Baddeley, Thomson, and Buchanan (1975) assumed that articulatory rehearsal is limited by what can be rehearsed within 2 s. Hence, it could be opposed that five letters are too many to remember with rehearsal alone. Future experiments could test the prediction more thoroughly.

3.4. Implications for theories of working memory in general

In Study 2, I observed the largest central attentional costs when people rehearsed aloud. As mentioned before, these costs were rather small. But then, in all other conditions the costs were even smaller or vanished completely. Still people were able to remember up to four words in their correct serial order in these conditions. This raises the possibility that information can be maintained in WM without paying attention to it. For example, in neural network models of WM (e.g., Farrell & Lewandowsky, 2002; Oberauer et al., 2012), information is maintained by binding content representations (e.g., a word) to context representations (e.g., a serial position). These models explain forgetting with interference between representations and usually do not assume additional rehearsal mechanisms to maintain information. Therefore, they offer a natural explanation for the clearly above chance memory performance when people neither used rehearsal, nor continuous refreshing or continuous elaboration to remember the words (i.e., the AS condition). Rerko and Oberauer (2013) showed in visual WM tasks that irrelevant information that could be forgotten was nevertheless maintained in WM. Importantly, this happened without focal attention and arguably there was no incentive for participants to rehearse, refresh, or elaborate this information. These results, together with the current results from the AS condition, question whether rehearsal mechanisms are especially important to explain WM capacity.

No theory of WM offers a straightforward explanation why central attentional demands of articulatory rehearsal increased disproportionately at set sizes 3 and 4. One explanation could be that people start to commit substantially more rehearsal errors from set size 3 on. In line with this explanation, it has been shown that rehearsal errors attract attention (Phaf & Wolters, 1993).

The discrepancy between the AS condition in my experiments and the AS condition in Vergauwe et al. (2014) stands out. Their average set-size slopes were more than three times larger than the ones I observed even though the used tasks were very similar. It remains largely unclear why there is such a difference in the slopes and this creates some ambiguity regarding the interpretation. One difference, however, is that my processing task required a visual-spatial decision (visual-spatial fit judgment) whereas their task required a verbal decision (parity judgment) in the AS condition. It could be that there was more domain-specific interference in their task than in my task. The whole difference is however unlikely attributable to domain-specific interference because they observed flat set-size slopes in the condition without AS, which should still lead to considerable domain-specific interference. Future research could examine this question more thoroughly by varying the domains of the memoranda and of the processing task independently. If the difference was not due to the domain of the processing task but due to noise, more data could still clear the picture how large the attentional costs in the AS condition actually are.

3.5. Implications for the understanding of central attention

What tells us Study 2 about the use of central attention in WM? I tested the assumption of the TBRS model (Barrouillet et al., 2004, 2007) that articulatory rehearsal can be carried out simultaneously with refreshing, an attentionally demanding rehearsal mechanism. The main conclusion is that rehearsing aloud a few words does require central attention. That is, other processes that demand central attention cannot be carried out without

any cost while people are rehearsing. A further conclusion is that elaboration only requires substantial central attention immediately after presentation of the memoranda, possibly for generating a mental image. After that, the delaying effect on processing was small, for example, because maintaining the image in WM is attentionally little demanding. Still, WM and delayed recognition performance tended to benefit from elaboration. This means that elaboration can be used as an efficient strategy to remember information in WM tasks.

More generally, the results of Study 2 showed that words can be maintained in WM without any rehearsal mechanism that requires central attention. But then, how is central attention used in WM if it is not required for WM maintenance? I suggest the same conclusion as in Study 1. Central attention serves the purpose to select representations for further cognitive processing or for physical actions.

4. Study 3: "Estimating Bayes Factors for Linear Models with Random Slopes on Continuous Predictors"

4.1. The necessity for a new statistics package

Recently, analyzing data in the Bayesian framework and with mixed-effects models has been recommended by statisticians. First, analyzing data in the Bayesian framework overcomes some of the problems associated with the classical statistical framework that relies on null hypothesis significance testing (NHST, Wagenmakers, 2007). With Bayesian statistics it is possible to compute the relative plausibility of two competing models with the Bayes factor (BF, Kass & Raftery, 1995). In other words, the BF tells us how likely one hypothesis is compared to another hypothesis given the data when the two hypotheses are equally likely a priori. This also means that we can obtain the evidence in favor of the absence of an effect (i.e., a null model/hypothesis), which is not possible with NHST. Furthermore, with Bayesian statistics it is possible to inspect the data during data collection and to decide whether more data are required or not, which is not recommended for NHST (i.e., optional stopping, Rouder, 2014). Finally, researchers are often confused by the different meanings of alpha levels and p values in NHST (Hubbard, 2004) and sometimes interpret p values as Bayesian posterior probabilities, which should not be done (Gigerenzer, 1993).

Second, with mixed-effects modeling (Bates, Mächler, Bolker, & Walker, 2014) it is possible to account for individual differences in the effect size of predictor variables (i.e., random effects). This is a departure from classical linear regression, in which each unit of observation (e.g., a participant, an item, or a study) is assigned the same effect size for a given predictor (i.e., a fixed effect). In the seventies, Clark (1973) showed that models that fail to include necessary random effects do not generalize well to an underlying population. Furthermore, Barr, Levy, Scheepers, and Tily (2013) showed that not accounting for random effects, which could potentially be present, and not accounting for correlations between random effects leads to biased inference.

The BayesFactor package (Morey & Rouder, 2014; Rouder, Morey, Speckman, & Province, 2012) for the R statistical environment (R Core Team, 2017) made a big step forward by combining Bayesian statistics with mixed-effects modeling. It provides users with default BFs and posterior distributions for the analysis of common experimental designs. It allows users to specify random slopes for categorical variables (e.g., variables with two, three, or more levels on a nominal scale) but not for continuous predictors that vary according to an ordinal or interval scale of measurement. Nevertheless, researchers often have variables in their data that are continuous and there may be individual differences in their effect size. That was exactly the problem I faced in Study 1 and in Study 2. For example, in Study 2 the data showed that processing RTs could be described well by assuming a linear relationship between memory set size and processing RTs. Therefore, I wanted to add set size as a continuous predictor into the models. Furthermore, model comparison revealed that set size had to be added as a random effect into the models. This type of analysis was not possible with previously available statistics packages. Therefore, I developed an own R package called BayesRS that is able to compute default BFs for categorical and continuous predictors with associated random slopes. The package is available in R and can be downloaded on CRAN under https://cran.r-project.org/web/packages/BayesRS/index.html.

4.2. Summary of Study 3

In Study 3, I tested the functionality of BayesRS in five simulation studies. The main goal was to make sure that the way the BF is computed in BayesRS yields reliable and valid results. The second goal was to test how different parameterizations of the random-effects structure affect the evidence in favor of an effect in terms of the BF. Kass and Raftery (1995) give a rough descriptive statement how to interpret BFs: BFs from 1-3.2 are not worth more than a bare mention. BFs from 3.2-10 are regarded as substantial, BFs from 10-100 as strong, and BFs that are larger than 100 as decisive evidence. The results show that BayesRS is able

to compute BFs reliably, producing trustworthy results. Crucially, the results show the same tendency as found when using p values for inference (Barr et al., 2013): Not accounting for actually present random slopes massively overestimates the evidence in favor of an effect. Finally, accounting for correlations between random effects underestimates the BF, especially when sample size is small.

The results from the first three simulation studies can be summarized as follows. Reliability and validity of the BFs computed with BayesRS are good as long as the true BF for the alternative hypothesis, as estimated with a benchmark method, does not grow large. BFs around 10, which are crucial according to the classification by Kass and Raftery (1995) to decide whether the evidence for an effect is substantial or strong, are estimated with high reliability. When the true BF grows larger (approx. > 1000) BayesRS starts to overestimate it systematically. This is, however, not particularly worrisome, because it usually does not matter for inference whether the BF is 1000 or 1400. In any case, it signals strong evidence in favor of the alternative hypothesis.

The fourth simulation study showed that failing to include random slopes that are present in the data leads to biased inference. Most problematic is that a model without random slopes massively overestimates the evidence in favor of the Alternative when the Null is actually true. In this scenario, it signaled strong evidence in favor of the Alternative in 42% of the cases. In contrast, a model including random slopes led to the same conclusion in only 1% of the cases. Therefore, including random slopes in a model is of great importance, also in the Bayesian framework. The fifth simulation study showed that the BF is underestimated when a model accounts for correlations between random effects. This was both the case when the Null and the Alternative were true. Therefore, modeling the correlations is not recommended when the focus of a study is on obtaining the evidence for the group/fixed effects. However, modeling correlations between random effects may still be insightful, especially when the correlations themselves are of primary interest in a study.

PART II: STUDIES

5. How Does Chunking Help Working Memory?

Mirko Thalmann, Alessandra Souza, and Klaus Oberauer

University of Zurich

Submission status:

Accepted for publication in: *Journal of Experimental Psychology: Learning, Memory, and Cognition*

Authors‘ contributions:

Mirko Thalmann: Development of the research questions, programming the experiments, analyzing the data, writing the manuscript

Alessandra Souza: Co-supervising the project, discussing the research questions and the results, commenting on the manuscript

Klaus Oberauer: Supervising the project, discussing the research questions and results, commenting on the manuscript

Acknowledgement

This research was supported by a grant from the Swiss National Science Foundation to KO (#149193).

5.1. Abstract

Chunking is the recoding of smaller units of information into larger, familiar units. Chunking is often assumed to help bypassing the limited capacity of working memory (WM). We investigate how chunks are used in WM tasks, addressing three questions: (1) Does chunking reduce the load on WM? Across four experiments chunking benefits were found not only for recall of the chunked but also of other not-chunked information concurrently held in WM, supporting the assumption that chunking reduces load. (2) Is the chunking benefit independent of chunk size? The chunking benefit was independent of chunk size only if the chunks were composed of unique elements, so that each chunk could be replaced by its first element (Experiment 1), but not when several chunks consisted of overlapping sets of elements, disabling this replacement strategy (Experiments 2 and 3). The chunk-size effect is not due to differences in rehearsal duration as it persisted when participants were required to perform articulatory suppression (Experiment 3). Hence, WM capacity is not limited to a fixed number of chunks regardless of their size. (3) Does the chunking benefit depend on the serial position of the chunk? Chunks in early list positions improved recall of other, not-chunked material, but chunks at the end of the list did not. We conclude that a chunk reduces the load on WM via retrieval of a compact chunk representation from long-term memory that replaces the representations of individual elements of the chunk. This frees up capacity for subsequently encoded material.

The scripts for all the experimental procedures and the raw data of the present article are available on the following Open-Science Framework webpage: https://osf.io/jjfbh/?view_only=3ebcbef89c3545019f6fde0fe28729f3.

Keywords: Short-Term Memory, Working Memory, Long-Term Memory, Chunk

5.2. Introduction

When people are required to remember a number of items over a brief retention interval they usually cannot recall more than a few of them (Brener, 1940). For example, when presented with a random sequence of geometric figures, participants can only recall about five or six of them in correct order. This limitation in remembering information over the short term is due to a limited-capacity store called working memory (WM). Although WM is severely limited in its capacity, some people are able to remember lists far exceeding this capacity. For example, when chess players at beginner level are briefly presented with middle-game positions in a chess game, they can recall about four pieces in the correct location (Chase & Simon, 1973). In contrast, players at advanced level or at master level can recall about eight or 16 pieces correctly, respectively. If WM is capacity limited, how can there be such a large difference in memory for chess positions as a function of chess expertise? Chase and Simon assumed that experienced chess players encode the positions as larger perceptual chunks, "each consisting of a familiar sub configuration of pieces" (p. 80). This interpretation is supported by the fact that, when presented with random constellations of chess pieces (i.e., shuffling the pieces of middle games), the number of pieces remembered by master players dropped to the level of beginners. It follows that pre-existing knowledge in terms of chunks can boost immediate memory. The present study focuses on this increase in immediate memory and asks in more detail what processes contribute to it.

The process of chunking was first described by Miller (1956) as the recoding of several presented stimuli into a single familiar unit or chunk. Miller proposed that chunking is achieved by grouping or organizing a sequence of inputs, and recoding it with a concise name. Therefore, remembering just the name essentially reduces the storage load on WM, arguably freeing capacity for storage of additional information. The reference to "familiar" units can be understood here as referring to the reliance on long-term memory (LTM) representations. In a similar vein, Cowan (2001) defines chunks as groups of items that have strong, pre-existing

associations to each other but weak associations to other items. To summarize, both authors highlight the importance of LTM in their definition of a chunk. The present work sought to make a step towards a deeper understanding of how people make use of chunks in WM tasks. In particular, we were interested in the question how chunks can reduce the load on WM.

Following Miller, several theorists have assumed that WM capacity is limited in terms of the number of chunks. Most prominently, Cowan (2001) proposed that about four chunks are available at a certain point in time in the focus of attention. Other researchers have also embraced the chunking idea, although they have different views on the exact number of chunks that can be held in WM (e.g., Chase & Simon, 1973; Gobet & Clarkson, 2004).

Evidence for the fixed-chunk hypothesis comes from experiments that varied chunk size and observed an approximately equal number of chunks recalled across different chunk sizes (Chen & Cowan, 2005, 2009; Cowan, Chen, & Rouder, 2004). In Chen and Cowan (2005), for example, participants were first trained to remember the pairings in a set of word pairs. Training also involved a set of single words for which participants had to remember that they were not paired with another word. Training proceeded until recall was 100% accurate. At this point, each individual trained word (henceforth singleton) and each of the word pairs were considered a chunk. Next, participants attempted immediate serial recall of word lists ranging from 4 to 12 words. The lists consisted either of learned word pairs, learned singletons, or new words. When recall of words was scored regardless of order, participants recalled approximately 3 chunks across all conditions (i.e., twice as many words from lists of word pairs than from lists of singletons or new words), consistent with the assumption that WM capacity is limited to a fixed number of chunks. Chen and Cowan (2009) showed that the constant number of recalled chunks can be observed best when participants had to engage in concurrent articulation ("articulatory suppression", AS) during encoding. The authors

conclude that the fixed capacity limit on chunks is most directly reflected in performance when articulatory rehearsal of phonological representations is prevented.

These experiments leave open two possibilities of how chunks help immediate recall. First, chunks require less capacity and therefore free up capacity in WM. This would be the case, for example, if remembering a single word required the same capacity as remembering a learned pair, because all were remembered as one chunk (Cowan, 2001). This account posits that a chunk is represented in WM independently of its composite elements, for example by a concise name as suggested by Miller. In contrast, a second possibility is that information from LTM assists in the reconstruction of the complete chunk from partial information in WM (Hulme et al., 1991, 1995). This account assumes that learned pairs (chunks) are maintained in WM in the same way as random pairs. However, at recall there is more LTM knowledge available for a previously learned pair compared to a random pair to assist reconstruction of the original set of elements. To distinguish these two possibilities, one needs to test memory for other items in the presence of a chunk: If chunking reduces the load on WM capacity, the presence of a chunk in a memory list should improve memory for other items maintained concurrently. In contrast, if chunks benefit only from being more successfully reconstructed at retrieval, other information in WM should not inherit that benefit.

To the best of our knowledge, only the study by Portrat, Guida, Phénix, and Lemaire (2016) presented mixed lists consisting of chunks and not-chunked items. However, they did not assess whether chunks improved the retention of not-chunked items. Therefore, it is still an open question whether chunks actually reduce the load on WM.

The experiments showing evidence that WM capacity is limited by a fixed number of chunks (Chen & Cowan, 2005, 2009; Cowan et al., 2004) had an additional feature, which makes their interpretation difficult: Each pair to be remembered consisted of unique elements. That is, each word could only occur as a single word or as a member of one pair. This is not

generally the case with chunks. Consider the often cited examples FBI and CIA, or any other acronym – they consist of letters that also occur in many other acronyms, and when known acronyms are included in lists of letters presented for short-term retention, each letter that figures as an element of an acronym can also occur as a singleton in the list (Portrat et al., 2016).

The use of chunks with unique elements in the experiments of Chen and Cowan has two important consequences. First, chunks can be detected instantaneously after presentation of the first word. This makes the encoding of the second word unnecessary, because participants already know the second word. Second, participants only have to remember the first word of a pair to remember both words. Even though doing so reduces the load on WM for this special kind of chunks, it is unclear whether this holds for chunks in general. Together, these two consequences of the unique-element chunks in the Chen and Cowan studies are sufficient to explain why participants remembered a fixed number of chunks: If for each learned pair they only encoded the first word into WM, then they encoded the same number of words into WM in all experimental conditions. This would be so regardless of whether or not WM capacity is limited to a fixed number of chunks. To circumvent this problem, in the current investigation we also investigated the chunking benefit with chunks consisting of not-unique elements.

5.2.1. The Present Study

The aims of the present study were three-fold. Our first goal was to assess whether chunking information in WM frees capacity to hold other, not-chunked information. To test for this possibility one needs to assess the impact of the chunk on the recall of other not-chunked information in WM. This prediction is best explained by means of an example. Assume that two lists have to be remembered: List 1 = F-B-I-D-Q-B, and List 2 = I-F-B-D-Q-

B. In List 1, the first three items form a single chunk (i.e., F-B-I). If encoding these three letters as a chunk reduces the load on WM from 6 items to 4 items, then there will be more free capacity to hold the second half of the list (i.e., D-Q-B) in List 1 than in List 2. Consequently, short-term retention of the second half of the list should be better in List 1 than in List 2.

The second aim of the present study was to provide yet a stronger test of the tenet of Cowan (2001) that the chunking benefit is independent of chunk size because chunks are assumed to be the basic storage units in WM. To do so, we compared recall of not-chunked lists while participants had to also hold in mind a chunk varying in size (e.g., 2-item vs. 4-item chunks). If capacity is independent of chunk size, then the chunking benefit for the not-chunked lists should be of similar magnitude when participants hold a smaller or larger chunk in WM.

Third, we examined for the first time how the chunking benefit is moderated by the requirements of detecting the chunk while at the same time holding other information in WM. This point is particularly important in relation to the type of material (i.e., chunks of unique vs. not-unique elements). When chunks consist of not-unique elements, participants need to encode all individual items before they can detect a chunk. This implies that at least temporarily their WM is loaded with the individual elements of the chunk, before these elements can be replaced by a single representation of the chunk. The temporary encoding of multiple elements could already damage other information in WM (i.e., through interference, or competition for rehearsal) before the WM load is reduced through chunking. Moreover, the reduction of WM load requires removing the individual elements from WM, so that the chunking benefit also depends on the efficiency of this process. Therefore, the chunking benefit for not-chunked information in WM might be much reduced, or even eliminated, when chunks consist of not-unique elements.

If the elements of a chunk must initially be encoded into WM individually before they are replaced by a more compact chunk, then we should also observe that the chunking benefit varies as a function of when a chunk is presented within a sequentially presented memory set. When presented at the beginning of the set, there is no other information in WM that could suffer from the temporary presence of the individual elements in WM, and these elements can be removed efficiently through a complete wipe-out of WM before encoding the compact chunk (Ecker et al., 2014). When the chunk appears later in a trial, individual items start to interfere with earlier encoded information before they can be replaced by the chunk representation, and that replacement involves targeted removal of only the items that belong to the chunk. The selective removal of individual elements from a memory set in WM is a more difficult process than the wholesale removal of the entire set (Ecker et al., 2014). As a consequence, the chunking benefit for simultaneously maintained not-chunked information should be smaller the later a chunk is presented as part of a memory set.

We tested the predictions detailed above with four experiments. In Experiment 1, we trained participants to recall chunked lists consisting of 2 or 4 words. Each chunk consisted of unique words. Next, we asked participants to hold two lists of 2 or 4 words for an immediate serial recall test. Each list was either a randomly arranged set of words (new lists) or a chunked list. Our main interest was in assessing whether recall of new lists was better in the presence of chunked lists (hereafter referred to as a chunking benefit) and whether this benefit was independent of chunk size. In Experiment 2, we tested the chunking benefit in conditions in which the status of the list (chunk or new) was unknown until all items of the list were presented, because the same item could occur in a chunk and in a not-chunked list. Moreover, we tested whether the load on WM from a chunked list was similar to that from holding a single-item representation (singleton) in WM. We also tested that question with AS (Experiment 3) to see whether any difference between large and small chunks could be attributed to differences in the duration of articulatory rehearsal. Finally, Experiments 2-4

investigated whether the chunking benefit depends on the serial position of the chunk within a trial.

5.3. Experiment 1

In Experiment 1, we tested the prediction that including a chunk in a memory set reduces the load on WM, thereby improving recall of other, not-chunked information maintained simultaneously with the chunk. To test this prediction, participants were presented with a memory set consisting of two short lists, followed by serial recall tests of each list. Each list was either a chunk or a new list composed of singleton words. Chunked lists were learned by heart in a training phase. As in the studies by Chen and Cowan (2005, 2009), each word could only occur either in a chunked list or a new list. If chunking frees WM capacity, then not only recall of the chunked list but also recall of the other not-chunked list on that trial should be better, compared to the condition in which both lists were new. Furthermore, we tested whether the chunk benefit was independent of chunk size. To this end we varied the length of the two lists independently of each other – new and chunked lists could comprise 2 or 4 items. If WM capacity is constrained by the number of chunks, but not by the size of each chunk, the benefit yielded by the chunk should be of similar magnitude irrespectively of the size of the chunked list.

5.3.1. Method

Participants. Twenty university students (15 women; M ≈ 25 years old) took part in two 1-hour sessions. They were compensated for participation with 30 Swiss Francs or partial course credit. All participants of the experiments reported in the present paper were native speakers of German. They were required to read and sign an informed consent form before the experiment started. In the end of the experiment, they were debriefed in detail about the purpose of the study.

Materials and Procedure. All experiments were programmed and run with the Psychophysics Toolbox 3 (Brainard, 1997; Pelli, 1997) in MATLAB. Participants sat at a distance of approximately 50 cm from the computer screen (viewing distance unconstrained). They were tested in individual cabins.

We constructed a pool of 338 one- and two syllabic nouns. None of the nouns started with the same three letters in the same order. For every participant, 24 nouns were randomly selected from this pool. Half of the nouns were used as singletons to construct new lists from. The other 12 nouns were used to construct four chunked lists: two chunked lists with 2 nouns; and 2 chunked lists with four nouns. Together, the singletons and the chunked lists were used as 16 sets to be learned in a training phase at the beginning of the experiment.

Training phase. In every training cycle, the elements of the 16 sets were displayed one by one across a row of black frames in the upper part of the screen. Set presentation was self-paced: participants started the presentation of each set in a cycle by pressing the spacebar. Words in sets of one (singletons), two, or four (chunked lists) were presented from left to right, each for 1000 ms. Order of presentation of the 16 sets was randomized in every cycle. After presentation of the 16 sets, cued recall tests of all sets followed. On every test, the first word of a set (randomly selected from all sets without replacement) was presented as a probe for 1000 ms in the top half of the screen, followed by presentation of four response boxes in

the bottom half. Presentation of the probe prompted participants to type all words belonging to that set, beginning with the first word (i.e., the word presented as probe). Participants could use the backspace key to correct any typing errors. They confirmed an answer by pressing the enter key. In case the tested set was a singleton or a 2-item chunked list, the remaining recall possibilities had to be skipped using the enter key. An answer was counted as correct when the first three letters of the word matched the word at that position in the set. Upper and lower case did not matter for responses to be counted as correct.

Participants completed a minimum of eight training cycles. Further training cycles were added until recall of all sets was 100% correct. Next, participants completed a distraction task. In the distraction task, participants had to judge the accuracy of 40 multiplication equations consisting of two factors in the range of 3-9. About half of the equations were correct. Participants pressed the left or right arrow key to indicate whether the displayed result of the multiplication was correct or incorrect, respectively. After the distraction task, we tested memory for the sets again via probed recall to ascertain that the chunks had been learned in LTM. As long as recall was not perfect, additional cycles consisting of set presentation, probed recall, distraction, and again probed recall to test LTM were added. Only when all 16 word sets were recalled correctly in the LTM test, participants proceeded to the main experimental phase.

Main experimental phase. Every trial started with the presentation of a fixation cross in the middle of the screen for 500 ms. Then, two lists were presented sequentially, one in the upper and one in the lower part of the screen. Nouns within each list were presented sequentially for 1000 ms across a row of black frames. Presentation of the last noun of the first list was followed immediately by presentation of the first noun of the second list without any free time in between. The two lists were probed in random order, 500 ms after list

presentation. Recall within a list proceeded in forward order. Participants typed the words using the keyboard. The same scoring as in the Training phase was applied.

We independently varied the size (2 or 4 items) and type (new or chunk) of list 1 and list 2, as well as their order of recall, resulting in 32 conditions. We aimed at replicating each condition eight times, and presenting them in random order across trials. However, due to a programming error, the 32 conditions were replicated 128 times resulting in 4096 trials that were randomized. The first 256 trials of this set were presented to the participants, resulting in an unbalanced design. In our analysis, we collapsed data across order of list presentation (list 1 and list 2) to increase the average number of data points per design cell available to be analyzed per participant (M = 16.00, SD = 3.76).

5.3.2. Results

Data Analysis. Frequentist statistics that are commonly used in psychological research have several shortcomings. For example, p-values express how likely the data are, given that the null hypothesis is true. However, researchers are usually interested in the reverse direction of inference: How likely is a hypothesis given the data? To overcome this shortcoming, and some other shortcomings associated with frequentist statistics (e.g., Wagenmakers, 2007) we used Bayesian statistics for all analyses reported in the present article.

In Bayesian statistics, the believability of model parameter values – such as the effect size of an experimental manipulation – is expressed as prior distributions (hereafter priors). The priors are updated with the likelihood of the data to yield posterior distributions of the parameters. Therefore, inference based on posteriors combines all the available information about the model parameters (e.g., Kruschke, 2014). The 95% highest-density interval (95% HDI) of the posterior covers the range of parameter values that are 95% credible after the data have been observed. Hence, the 95% HDI can be used to inform about uncertainty of

parameters in question. In the present work, we do not display classical confidence intervals but use HDIs of the posteriors instead, because they can be interpreted straightforwardly, which is not the case for confidence intervals (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016). For descriptive purposes, we also plot the standard error of the mean to show the variability of the data without the assumptions of any model.

In Bayesian statistics, models can be compared using the Bayes Factor (BF). The BF quantifies the strength of evidence in favor of one model in comparison to another, competing model, given the data. For example, we can compare a model including an effect in an ANOVA (e.g., a main effect, a two-way interaction) to a model omitting this effect. A BF in favor of the former model reflects the evidence in favor of the effect; the inverse of the BF states the evidence in favor of the Null hypothesis that the effect is absent. As a rough guideline to interpret the quantity of BFs, Kass and Raftery (1995) suggest that BFs between 1 and 3.2 are not worth more than a bare mention, BFs between 3.2 and 10 are substantial evidence, BFs between 10 and 100 represent strong evidence, and BFs > 100 are seen as decisive. In the present article, we computed the BF for t-tests with the BayesFactor package (Morey & Rouder, 2014) using the default priors. All other BFs were computed with self-constructed JAGS (Plummer, 2003) models using the Savage-Dickey density ratio, which provides the BF for nested models (Lee & Wagenmakers, 2014; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009). When comparing a model in which the parameter of interest is allowed to vary freely to a model in which the parameter of interest is fixed to the value of the Null model (usually zero), the BF can be obtained by dividing the height of the prior by the height of the posterior at the parameter value of the Null model.

Whenever we used self-constructed JAGS models, we applied a two-stage procedure to determine the BFs of the effects of interest. First, we selected the best-fitting model out of a set of three models according to the DIC[1], which can be used for model selection of

[1] The DIC relates to other information criteria as it uses a classical estimate of fit and penalizes for the effective number of parameters in a model.

hierarchical Bayesian models (Spiegelhalter, Best, Carlin, & van der Linde, 2002). The three fitted models varied in their hierarchical structure. Model 1 assumed a random intercept for each subject but only fixed effects on all other parameters. Model 2 assumed random effects on all parameters in the model except for the highest-order interaction. Model 3 assumed random effects on all main effect parameters, but no random effects on interactions. Second, we computed the BFs for all effects of interest within the winning model with the Savage-Dickey density ratio. In the following, we only report the BFs of the winning model.Interested readers are referred to the supplementary material if they want to see which model won the DIC comparison in each case.

Serial Recall of New Lists. Our main interest was in how the accuracy of recall of new lists is influenced by the presence of a chunk, and by the size of that chunk, compared to trials in which both lists were new. If storing a chunk frees WM capacity, we should observe a main effect of list type, with better recall of new lists when the other list is a chunk than when it is another new list. If chunks impose a constant load on WM regardless of their size, then recall of a new list should not vary as function of chunk size, whereas recall would decrease when the other list is a new list with size 4 compared to size 2, resulting in an interaction between other-list type (new or chunk) and other-list size (2 or 4 items). The relevant data are displayed in Figure 1 and in Figure 2. Note that recall of new lists with 2 and 4 items are presented in different subpanels in Figure 1.

We analyzed serial recall accuracy of new lists with two Bayesian linear regressions (i.e., one for each tested set size[2]). We entered the effects for type of the other list (new list or chunk), size of the other list (2 or 4 items), time of recall of the new list to be analyzed (first vs. second), and all higher-order interactions between these predictors.

[2] The reason for running separate analyses is that we wanted to obtain the evidence in favor of the disordinal Size Other x Type Other interaction (left panel in Figure 1). Evidence for a disordinal or cross-over interaction is more convincing than for an ordinal interaction because a disordinal interaction remains when a non-linear transformation is applied to the dependent variable (Loftus, 1978).

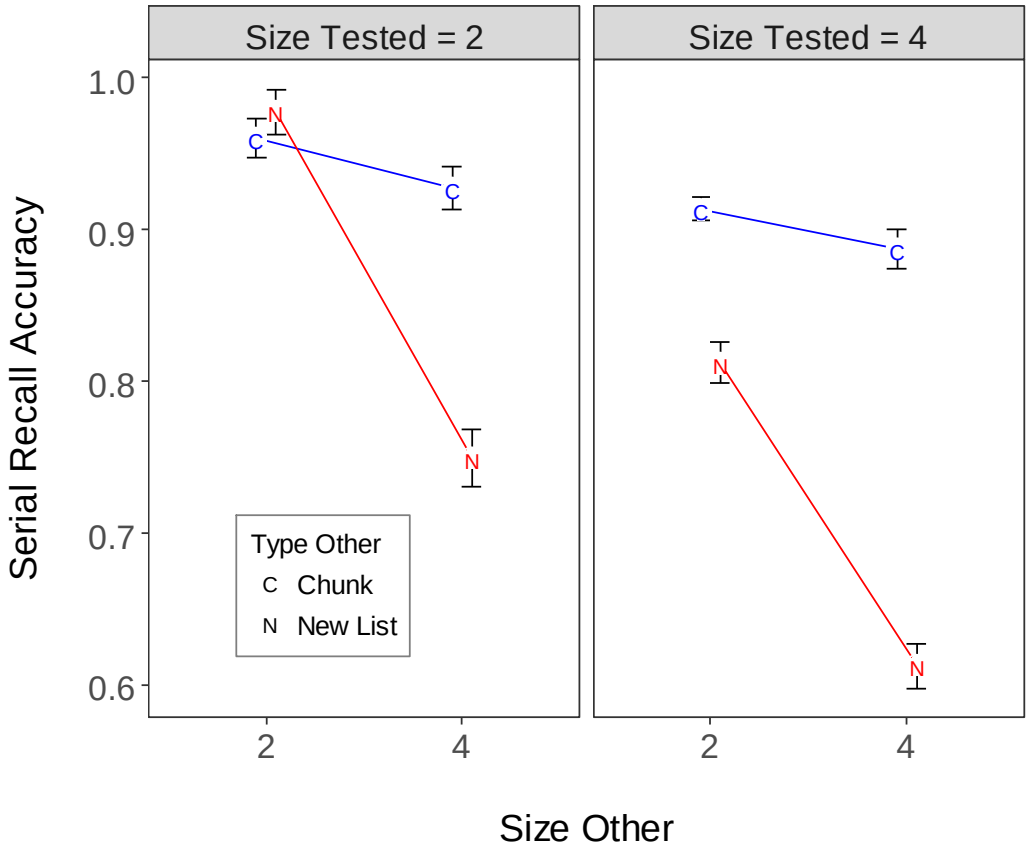

Figure 1. Serial recall accuracy (proportion correct) of new lists in Experiment 1. Recall of new lists with 2 and 4 items are presented in different sub-panels. Performance is plotted as a function of the size of the other list (2 or 4 items) for the two types of other lists (new list or chunk). Error bars represent within-subjects standard errors.

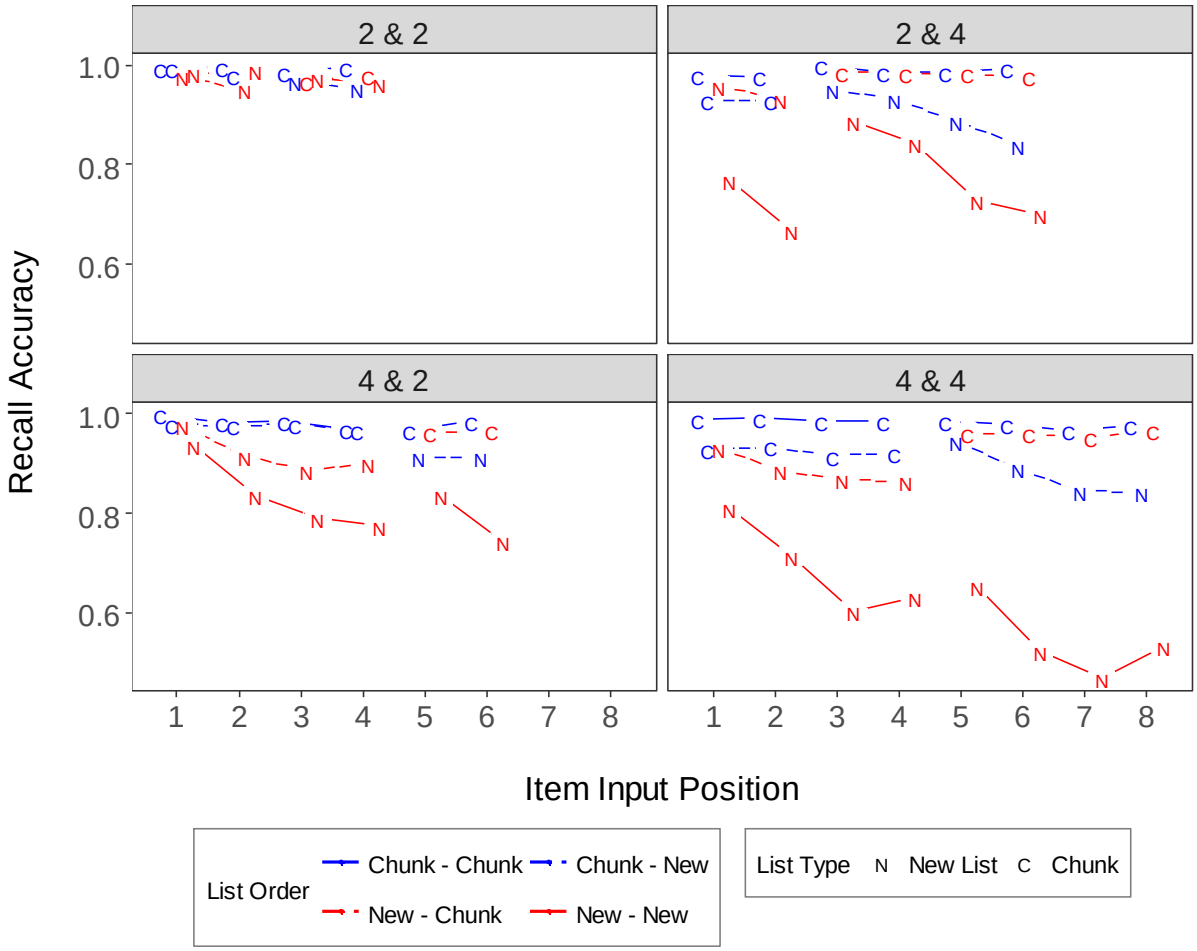

Figure 2. Serial recall accuracy (proportion correct) in all 16 conditions of Experiment 1. Note. For better interpretability we omitted error bars.

Table 1.

Posterior Means, Lower and Upper Boundaries of the 95% HDIs, and the BFs for the Parameters of the Linear Model fitted to the Data of Experiment 1.

		Measure			
			95% HDI		
	Effect	Posterior Mean	Lower Bound	Upper Bound	Bayes Factor
Set Size = 2					
	Type Other	0.08	0.05	0.11	249222
	Size Other	-0.13	-0.16	-0.10	3.90E+16
	Time of Recall	-0.11	-0.13	-0.08	7.50E+10
	Type Other x Size Other	0.19	0.14	0.25	1.10E+09
	Type Other x Time of Recall	0.06	0.01	0.12	4.60E-01
	Size Other x Time of Recall	-0.16	-0.21	-0.10	5.81E+05
	Three Way	0.19	0.08	0.30	2.90E+01
Set Size = 4					
	Type Other	0.18	0.16	0.21	1.30E+36
	Size Other	-0.11	-0.14	-0.08	1.20E+11
	Time of Recall	-0.11	-0.13	-0.08	6.20E+10
	Type Other x Size Other	0.17	0.12	0.22	3.72E+06
	Type Other x Time of Recall	0.09	0.03	0.14	5.50E+00
	Size Other x Time of Recall	-0.07	-0.12	-0.01	7.30E-01
	Three Way	0.08	-0.03	0.19	0.24

The statistical evidence for the fixed effects in the two separate regression models is shown in Table 1. There was decisive evidence in favor of the main effect of other-list type, which means that a new list was remembered better when it was presented together with a chunk compared to another new list. This finding supports the hypothesis that chunks reduce the load on WM. The evidence was also decisive in favor of the interaction between other-list type and other-list size. This indicates that increasing the size of a chunk leads to a smaller additional load on WM than increasing the size of a new list. Whether there is any effect of chunk size at all was evaluated in an additional analysis reported below.

The other reported main effects show that new lists are remembered worse in the presence of long than short lists (main effect of other-list size), and that new lists were recalled worse when probed second rather than first (main effect of time of recall). Worse recall of lists probed second is evidence for output interference from the first-recalled list (e.g., Cowan, Saults, Elliott, & Moreno, 2002; Jones & Oberauer, 2013; Oberauer, 2003). The results were inconclusive on whether output interference is stronger (a) from a new list than from a chunk (Type Other x Time of Recall) and (b) from a longer list than from a shorter list (Size Other x Time of Recall) as the BFs in the two parallel analyses differed slightly (see Table 1).

We next ask whether the size of a chunk affects how well a new list is recalled. This analysis tests the hypothesis that a chunk imposes the same load on WM regardless of its size. If so, chunk size should have no effect on how well a new list can be maintained concurrently. However, chunk size could matter for output interference because larger chunks require output of more words. Our previous analysis yielded inconclusive evidence for output interference. Nevertheless, to exclude any potential effects of output interference, for this analysis, we zoomed in on trials in which the new list under consideration was recalled first, thus removing any potential contribution of output interference. We compared recall accuracy of new lists in the conditions in which the other list was a chunk, and chunk size was either 2 or 4. If chunks reduce the load on WM independently of their size, recall accuracy of new lists should not differ as a function of chunk size. The BFs supported this claim (41.7 and 16.4 in favor of the null for new lists of size 2 and 4, respectively), which is strong evidence for the assumption that chunk size has no influence on how much capacity a chunk requires in WM.

Serial Recall of Chunked Lists. Chunked lists were recalled with a very high level of accuracy (M = .974). Recall of chunked lists varied as a function of other-list type (BF = 535), and other-list size (BF = 612). Chunks were remembered better together with another chunk

(*M* = .985) than with a new list (*M* = .962), and better together with a list of size 2 (*M* = .985) compared to a list of size 4 (*M* = .962). There was no evidence for a main effect of size of the current chunk to be remembered (BF = 145 for the Null), and there was no substantial evidence for any higher-order interaction between these factors (all BFs < 3.2 for the Alternative).

5.3.3. Discussion

Experiment 1 used a design modeled after Chen and Cowan (2005) using a training phase to establish chunks of different sizes (2 or 4 words). It showed that storing a chunk reduced the load on WM: Remembering a random list of words was easier when another list to be remembered concurrently was a chunk than when it was another random list. Moreover, our data supported the claim that WM capacity is constrained by the number of chunks, and not by chunk size: Increasing the size of a chunk had no influence on remembering a random list of words when the random list was recalled first, whereas increasing the size of another random list did impair recall. The size of a chunk had no influence on how much WM capacity it requires, in line with the prediction of the embedded-process theory of Cowan (2001).

Chen and Cowan (2005) suggested that people may only hold the first item of a chunk in WM, which allows retrieval of the subsequent items from LTM. This possibility exists for their earlier chunking experiments as well as for Experiment 1, because in these experiments all chunks consisted of unique elements, which could not be an element of any other chunk or new list. This peculiarity of the design has two potentially important implications. First, participants can use the first word of a chunk as retrieval cue to recall the entire chunk. Second, the chunk can already be detected upon presentation of the first element, so that any further elements do not even have to be encoded into WM, thereby avoiding the risk that they

interfere with other information held in WM. Under these circumstances, any theory predicts that chunk size does not impact capacity, because what participants do is just remembering one word for chunks of any size.

We argue that chunks consisting of unique elements are not representative for naturally occurring chunks. In many situations, chunks do not start with unique elements. In these situations, it is not possible to stop encoding after the first item of the chunk has been presented. Further, remembering only the first item is not an advisable strategy either because other lists also start with the same item. Hence, application of the strategy to only remember the first item of a chunk is often not possible. To investigate the hypothesis that chunks reduce the load on WM irrespectively of chunk size, one has to use chunks with not-unique elements. This is what we did in the following experiments.

5.4. Experiment 2

In Experiment 2, we again tested whether there is a chunking benefit and whether it is independent of chunk size. However, we made sure that people cannot apply the simple strategy to only encode and remember the first item of a chunk. To that end we used chunks consisting of not-unique elements in Experiment 2.

We suspect that the chunking benefit may decrease when using not-unique elements that can also appear in other lists and other chunks. A chunk composed of not-unique elements cannot be detected before all its elements are encoded individually into WM. We explain the hypothesis by means of an example. Let us assume that a participant encodes a list in which the first two letters are F and B. The third letter can be either an I or a Q. The participant can only replace the individual representations of the letters with a chunk representation in the first case, in which the sequence is the well-known chunk FBI, but not in the second case, in which the sequence is FBQ. Replacement of the individual letters by a

chunk is not possible until presentation of the third letter. Before that, the encoding processes in both cases are the same.

Let us first focus on the case that the three-letter sequence is presented before some other items are presented. After presentation of all items of the chunk no representation of other WM contents will be degraded. If the chunk representation is independent of chunk size, it follows that successful replacement of the individual items with the chunk representation reduces the load on WM independent of chunk size. Now let us consider the case where the three-letter sequence is presented after some other information has already been stored in WM. The memory representations of the earlier presented items will be degraded and remembered less well after presentation of the chunk when all items of the chunk have to be encoded individually. This follows from any theory of WM, whether it ascribes the reason of forgetting in WM to decay with rehearsal counteracting decay, to a limited resource (as the theory of Cowan, in which the resource is a fixed number of discrete slots), or to interference: The representations of the earlier items have suffered more from decay, which could not be counteracted by rehearsal during encoding of the chunk, or they have received less of a share from the limited resource because it was redistributed to the individual chunk items, or they have suffered interference from encoding the individual items of the chunk. For these reasons, a chunk presented at the end of the entire memory set should be much less beneficial than a chunk presented at the beginning.

With regard to the above prediction about the serial position of a chunk, decay theories of WM differ subtly from resource and interference theories. A decay theory predicts a chunk benefit with a chunk in any position. This is because longer lists do not only take longer to be encoded, but also to be reactivated through rehearsal. A chunk at the end of the list does not reduce the amount of forgetting of other items during encoding, but it does reduce the amount of rehearsal needed. After detection of the chunk only the chunk representation has to be

rehearsed, which takes less time than rehearsing all individual representations. Hence, even though a decay theory predicts that chunks help more when presented earlier in the list, it additionally predicts that chunks in any position still help memory for not-chunked items because a chunk takes less time to be rehearsed. In contrast, resource or interference theories do not predict any benefit from a chunk presented at the end of the entire memory set.

To summarize, all theories predict that the reduction of WM load by replacing multiple items by a single chunk is beneficial for subsequently encoded material, because capacity is freed up only for material encoded after that replacement. A decay-and-rehearsal theory predicts additionally that chunks in any input position help memory for not-chunked items.

In Experiment 2, we assessed the effect of chunk size by comparing chunked lists of size 3 vs. singleton lists composed of a single letter. New lists in the present experiment consisted of 3 letters. As a consequence of this reduction of list size in comparison to Experiment 1, we had to increase the number of lists presented in every trial from two to three to circumvent potential ceiling effects. The training phase in the beginning of the experiment was omitted because we used well-known acronyms consisting of three consonants as chunks.

A prerequisite of Experiment 2 was that participants detected an acronym when presented in the very beginning of the list, but also when presented in the end of the list. Portrat et al. (2016) showed that the latter may not be guaranteed. They observed that acronyms were recalled worse when presented in the middle or in the end of a list compared to the beginning of the list in a complex-span paradigm. It could be that the chunks were not detected in these conditions because it was difficult to anticipate when a chunk began and when it ended. It is impossible to investigate the beneficial effect of chunks on not-chunked information when the chunks are not detected. To maximize the chance that participants

detected the chunks in any serial position of the memory set in the following two experiments, we broke the memory set into three clearly demarcated lists, each of which could be a chunk.

5.4.1. Method

Participants. Twenty university students (17 women; M ≈ 25 years old) participated in Experiment 2 for one session lasting approximately one hour. Participation was compensated with 15 Swiss Francs or partial course credit.

Materials. We constructed a pool of 30 known 3-consonant acronyms to serve as the chunked lists in Experiment 2. To create the 30 singleton lists, we took the first consonants of the chunked lists. To create the new lists of size 3, we shuffled the consonants of the chunked lists six times to construct 180 new lists. Shuffling sometimes re-created known acronyms. Therefore, individual consonants were exchanged with consonants from other new lists by hand to have all new lists differing from chunks. We deleted 15 new lists to obtain a total of 165 new lists because we needed 5.5 times as many new lists as chunks. The new lists were allowed to overlap with chunks in individual consonants at certain positions (i.e., position 1, 2, or 3) or in the beginning or in the ending pair (i.e., items 1 and 2 or items 2 and 3). To assess whether we were successful in creating chunked versus new lists, we compared those lists on two measures to assure that they only differed in overall familiarity but not in transition probabilities between consonants, which also influence short-term memory retention (Mayzner & Schoenberg, 1964). First, we compared the number of Google hits (restricted to Switzerland), which served as a measure of overall familiarity of the strings for the participants in the present experiments. If the chunks are more familiar, they should generate more Google hits than new lists. We computed $\log_{10}$ Google hits because of skew in the data and outliers in the upper range of the scale, and submitted them to a Bayesian t-test using the BayesFactor package (Morey & Rouder, 2014). The alternative hypothesis that

more hits were generated for chunks (M = 5.60) than for new lists (M = 4.19) was supported by a BF of 3.76×10^{41}. Second, we compared the case insensitive corpus frequencies (Heister et al., 2011, also $\log_{10}$ transformed) of the bigrams in the two types of lists with another t-test. A difference in this measure would suggest that some of the transitions between consonants are more frequent in one type of list. However, the BF was 2.1 for the null hypothesis, suggesting that chunks (M = 4.67) are not likely to differ from new lists (M = 4.47). Hence, our chunks as a whole were more familiar than the new lists but they were comparable in familiarity at the level of bigrams. All stimuli were presented twice across the experiment to control for familiarity within the experiment. The stimuli used in all experiments are available on the OSF webpage.

Procedure. In every trial, three rows, each consisting of three black box frames, were displayed on the screen. The rows were shown at the top, middle, and bottom of the screen. Each row served to present a memory list. Lists were presented in order from top to bottom. The letters composing each list were presented one-by-one (for 1 s each) from left to right. We were interested in the serial recall of new lists depending on the context they were presented in (together with a chunk or another new list), depending on chunk size (singleton lists vs. chunked list), and depending on the chunk position within a trial (beginning or the end of the trial). With this aim, we created five experimental conditions (see Figure 3). In all five conditions, two new lists of size three were presented. Only the remaining list varied between conditions. In the Singleton First and Singleton Last conditions, this critical list was a chunk of size one (singleton) presented in the top row or in the bottom row, respectively. In both conditions, the singleton was presented in the third serial position in a row, and it was preceded by two leading blanks. In the Chunk First and Chunk Last conditions, the critical list was a chunked list presented in the top row or in the bottom row, respectively. In the Baseline condition, all three lists were new lists of size three.

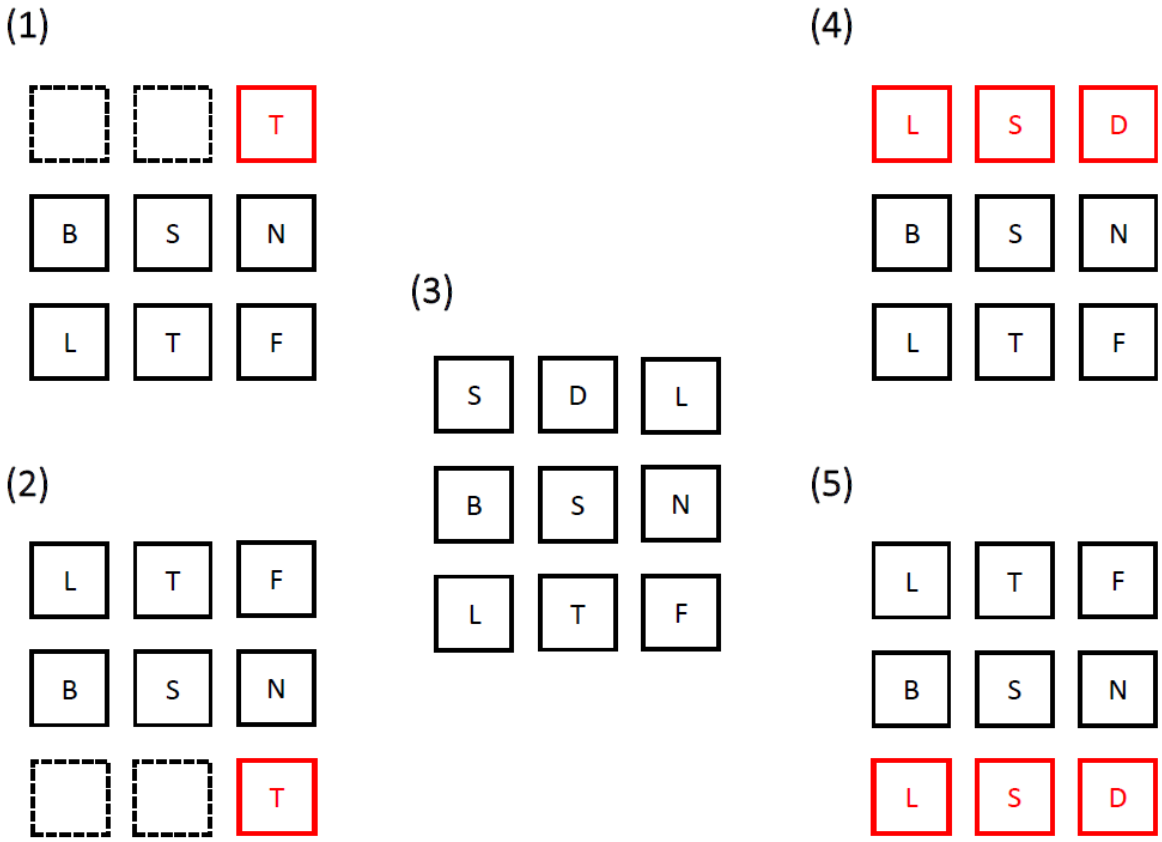

Figure 3. Example of the five experimental conditions in Experiment 2. Singleton lists (single letter) and chunked lists are identified in red (note that in the actual experiment all frames were black). The black dashed lines indicate frames in which no item was presented.

After presentation of all lists, the recall test started: Each row was cued to be recalled in left-to-right serial order. An empty black frame prompted participants to recall the list at the cued location (upper, middle, or lower row). For three-item lists or chunks, every consonant was sequentially cued with an individual empty frame. For singletons, only one empty frame at the respective third list position was presented as a cue. Participants confirmed responses with the enter key. In all five conditions, we controlled that each of the three rows was cued ten times as the first, ten times as the second, and ten times as the third list to be recalled. This resulted in 150 trials that were presented in a randomized order over the course of a single session.

5.4.2. Results

Serial Recall of Chunked Lists. As a manipulation check, we analyzed whether singletons and chunks were recalled better than new lists that were presented in comparable rows. Serial recall accuracies for all items in the five conditions are plotted in Figure 4 against item input position, and in Figure 5 for the three list types against row of presentation. It is clearly visible in both figures that singletons and chunks were recalled better than new lists. Surprisingly, three-element chunks tended to be recalled better than singletons. Across all input and output positions, the average recall accuracies were .74, .80, and .67 for singletons, chunks, and new lists, respectively. We computed pairwise comparisons of the means of the three list types in a Bayesian linearregression on proportion correct. There was overwhelming evidence for better recall of singletons vs. new lists (BF = 990), for better recall of chunks than new lists (BF = 2.9×10^{10}), but slight evidence against the apparent better recall of chunks than singletons (BF = 0.54).

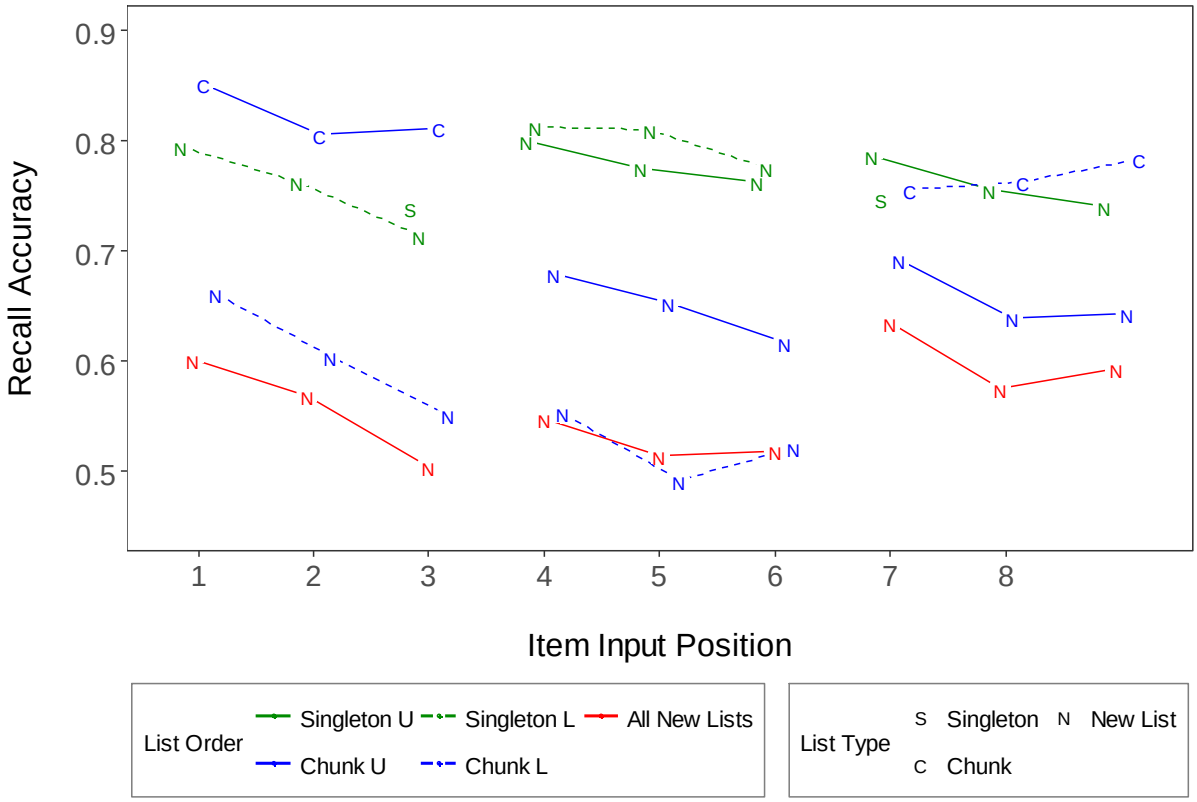

Figure 4. Serial recall accuracy (proportion correct) is plotted against item input position separately for each of the five experimental conditions in Experiment 2. For better interpretability of the figure, (a) we allocated singletons presented in the upper and lower row in the figure to item input position 3 and 7, respectively, and (b) we omitted error bars.

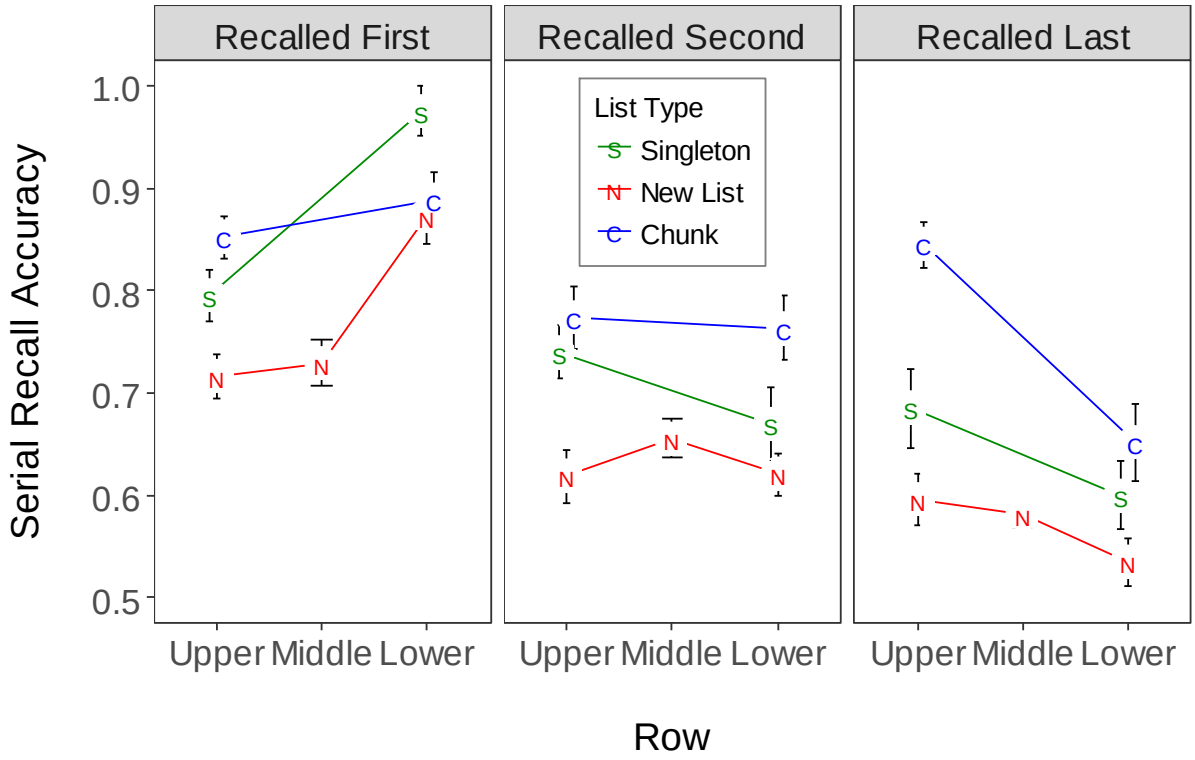

Figure 5. Serial recall accuracy of the three list types is plotted against row of presentation. The three panels represent the data from the three list output positions in Experiment 2.

Serial Recall of New Lists. Next, we focused on the impact of having a singleton or chunk in a trial upon recall of the other new lists. We performed two analyses – one on the data of the Singleton First and Chunk First conditions against the Baseline, and another one on the data of the Singleton Last and Chunk Last conditions against the Baseline. The first analysis focused on recall of new lists in in the middle and lower row, depending on whether a singleton, a chunk, or a new list was presented in the upper row in a trial. The second analysis focused on recall of new lists in the upper and middle row, depending on whether a singleton, a chunk, or a new list was presented in the lower row in a trial.

Chunk First. Mean serial recall accuracy of new lists when preceded by a singleton, a new list, or a chunk is shown in Figure 6a. We performed a Bayesian linear regression on these data with the variables row of presentation (middle vs. lower, zero-centered) and

condition. The analysis focused on the question whether a chunk that is presented in the beginning of a trial helps retention of new (not-chunked) lists presented afterwards. We coded the three levels of the condition variable in terms of two simple-code contrasts: baseline vs. singletons and baseline vs. chunks. Given that lists were probed in random order, there was a variable number of lists recalled prior to recall of a given list. To control for the output interference from these prior recalls, we added the number of previous list recall attempts (also centered on zero) as a control variable into the regression analysis. The variable was entered as a continuous covariate into the analysis because previous work showed that output position affects memory approximately linearly (Oberauer, 2003). In addition to that, we wanted to know whether output interference differed between list types (e.g., is output interference of a chunk the same as of a new list?). Therefore, we entered two further zero-centered predictors into the model. The first coded whether, in the chunk condition, the secondly recalled list was preceded by the recall of a new list or a chunk. The second coded whether, in the singleton condition, the secondly recalled list was preceded by the recall of a new list or a singleton.

Because some of the chunks may not have been familiar to all participants, we performed the same analysis only for those trials in which singletons and chunks were recalled correctly (Figure 6b). When chunks were recalled correctly, we can be more confident that participants actually recognized and remembered them as chunks, and also that there were no trade-offs between maintenance of the chunk or singleton and new lists. Posterior means, 95% HDIs, and BFs of the effects are shown in Table 2.

The evidence for the effects of interest was qualitatively the same across the two sets of analyses shown in Table 2. The evidence for better recall of new lists in the presence of a singleton or chunk was decisive. However, contrasting the effects for singleton and chunk showed that memory for new lists was better in the context of a singleton than in the context

of a chunk. There was no evidence for a main effect of row of presentation (middle vs. lower) on accuracy. The benefit of having a singleton in the upper row was larger in the middle row than in the lower row (Singleton x Row interaction), but the evidence was not strong. There was no evidence for the chunk benefit to differ between the middle and the lower row. As expected, the parameter for the number of previously recalled lists was negative. This shows that every additional list that has been recalled previously leads to worse memory for later recalled lists. Finally, there is moderate to strong evidence that output interference does not differ between different list types (Recalled Chunk Before and Recalled Singleton Before parameters).

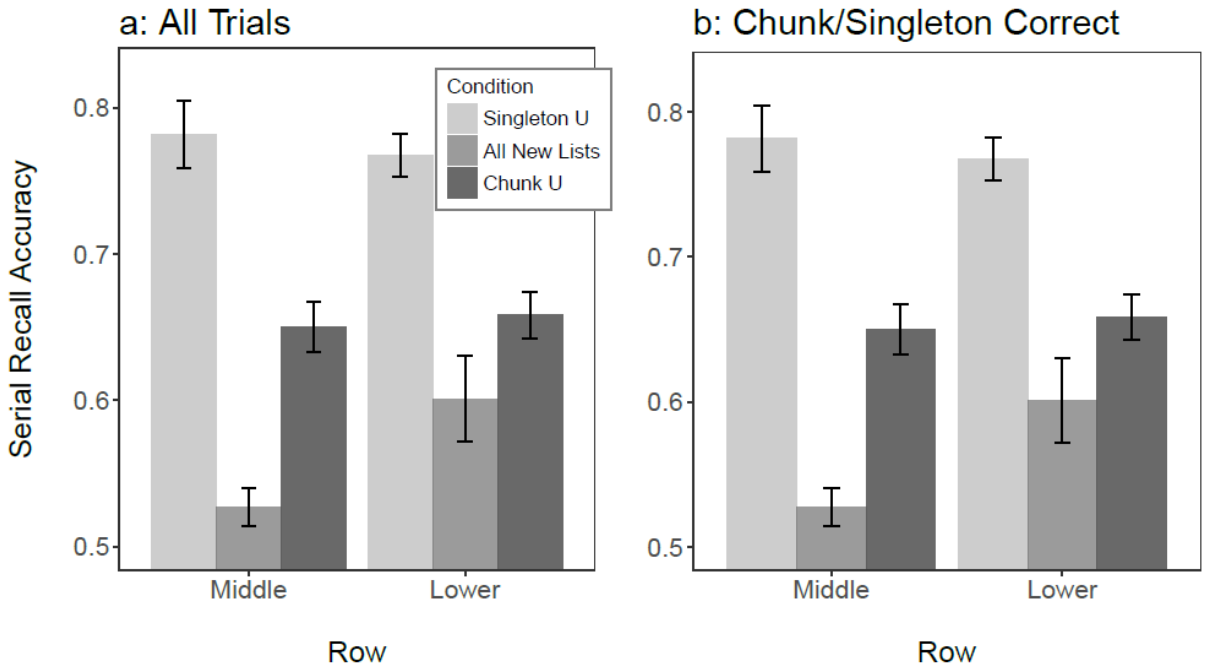

Figure 6. Mean serial recall accuracy of new lists that were preceded by a singleton, a new list, or a chunk plotted against row of presentation in Experiment 2. Panel a is based on all data, and Panel b is based only on those data from the Chunk first condition and the Singleton first condition in which participants recalled the chunk or the singleton correctly. The error bars represent within-subjects standard errors.

Table 2

Posteriors means, 95% HDIs, and the BFs of the parameters of the linear regression model fitted to the data of Experiment 2presented in panels a and b of Figure 6, respectively.

		Measure			
		Posterior Mean	95% HDI		Bayes Factor
Chunk First	**Effect**		Lower Bound	Upper Bound	
All Trials	Chunk (vs. Baseline)	0.09	0.06	0.12	187’566.00
	Singleton (vs. Baseline)	0.21	0.18	0.24	2.50E+36
	Row	0.02	0.00	0.05	0.10
	Chunk x Row	-0.06	-0.13	0.00	0.41
	Singleton x Row	-0.09	-0.15	-0.03	2.30
	Singleton vs. Chunk	0.12	0.09	0.15	2.70E+10
	Nr. Previous Recall Attempts	-0.12	-0.14	-0.11	2.00E+50
	Recalled Chunk Before	-0.03	-0.10	0.05	0.08
	Recalled Singleton Before	0.08	0.00	0.15	0.41
Chunks and Singletons Correct	Chunk (vs. Baseline)	0.10	0.06	0.13	92’066.00
	Singleton (vs. Baseline)	0.22	0.18	0.25	4.30E+30
	Row	0.02	0.00	0.05	0.09
	Chunk x Row	-0.06	-0.13	0.00	0.40
	Singleton x Row	-0.09	-0.15	-0.03	2.20
	Singleton vs Chunk	0.12	0.08	0.16	3.19E+05
	Nr. Previous Recall Attempts	-0.12	-0.14	-0.11	1.67E+50
	Recalled Chunk Before	0.05	-0.05	0.14	0.12
	Recalled Singleton Before	0.05	-0.05	0.14	0.12

Chunk Last. Next, we analyzed the data of the Singleton Last and Chunk Last conditions against the Baseline, which addresses the question whether a chunk helps retention of previously presented new lists. We ran the same analysis as for Chunk First (see Table 3). The data are shown in Figure 7. There was only evidence for a benefit for recall of new lists in the singleton condition, but not in the chunk condition. There was no difference in accuracy across conditions between the upper and the middle row. There was no evidence that having a singleton or a chunk in the lower row differentially affected memory for new lists in the upper or in the middle row (Singleton x Row and Chunk x Row interactions).

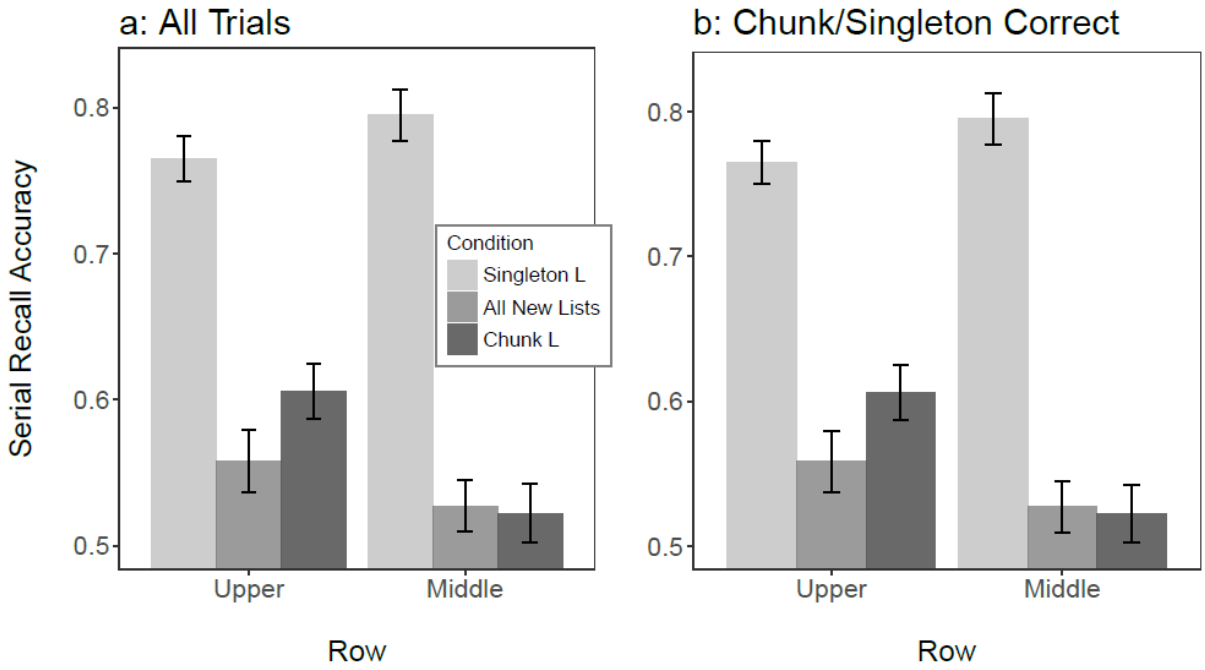

Figure 7. Mean serial recall accuracy of new lists when a singleton, a new list, or a chunk was presented in the lower row (Panel a: full data; panel b: Data from trials with correct recall of chunks/singletons) in Experiment 2. Error bars represent within-subjects standard errors.

Table 3

Posteriors means, 95% HDIs, and the BFs of the parameters of the linear regression model fitted to the data of Experiment 2 in panels a and b in Figure 7, respectively.

		Measure			
		Posterior Mean	95% HDI		Bayes Factor
Chunk Last	**Effect**		Lower Bound	Upper Bound	
All Trials	Chunk (vs. Baseline)	0.02	-0.01	0.05	0.06
	Singleton (vs. Baseline)	0.24	0.21	0.27	2.60E+50
	Row	0.03	0.00	0.05	0.25
	Chunk x Row	0.05	-0.01	0.11	0.22
	Singleton x Row	-0.06	-0.12	0.00	0.36
	Singleton vs. Chunk	0.22	0.19	0.25	2.4e+413
	Nr. Previous Recall Attempts	-0.07	-0.09	-0.06	3.60E+18
	Recalled Chunk Before	-0.03	-0.11	0.04	0.09
	Recalled Singleton Before	0.03	-0.04	0.10	0.08
Chunks and Singletons Correct	Chunk	0.02	-0.02	0.06	0.05
	Singleton	0.24	0.20	0.28	1.70E+34
	Row	0.03	-0.03	0.08	0.08
	Chunk x Row	0.05	0.00	0.11	0.25
	Singleton x Row	-0.06	-0.12	-0.01	0.47
	Singleton - Chunk	0.22	0.17	0.27	1.30E+15
	Nr. Previous Recall Attempts	-0.07	-0.09	-0.06	6.40E+13
	Recalled Chunk Before	-0.02	-0.10	0.07	0.08
	Recalled Singleton Before	0.01	-0.07	0.10	0.08

5.4.3. Discussion

Chunks were recalled better than random lists, demonstrating that our manipulation worked. Chunks even tended to be recalled better than singletons. The main finding of Experiment 2 was that chunks improved recall of later presented information but not recall of earlier presented information compared to random lists. This finding is consistent with theories explaining WM capacity by the allocation of a limited resource or by interference between representations encoded into WM. The fact that chunks in the lower row did not help recall of the preceding new lists (in the upper and middle row) at all is partially at odds with decay-and-rehearsal theories, as will be explained in the following.

All three theoretical perspectives assume that the Baseline condition and the Chunk Last condition do not differ until all items within a trial have been presented. The chunk itself can only be recognized after all items of the chunk have been presented (Bower & Springston, 1970). For instance, consider a resource theory with discrete resources (a.k.a. slots) sufficient to maintain four chunks (Awh, Barton, & Vogel, 2007; Cowan, 2001; Zhang & Luck, 2008). Immediately after presentation of the last item in the lower row, four items are held in four slots, and any additional items could not be accommodated. Even though the last three items can now be replaced by a single chunk, earlier presented items cannot be recovered because they dropped out of WM earlier. Therefore, memory for the first and second list should not benefit from recognition of a chunk in the lower row. In contrast, when the upper list is a chunk, a chunk representation can be retrieved from LTM, and the three individual item representations can be dropped from WM. As a consequence, participants have two more free slots in the Chunk First condition than in the Baseline condition that they can use to maintain subsequent items, resulting in better memory for the middle and lower lists.

An interference account (e.g., Oberauer & Lewandowsky, 2008) states that forgetting is due to items in WM interfering with each other. Right after seeing the last item, interference should be the same in the Chunk Last condition and in the Baseline condition because a chunk representation can only be retrieved from LTM after all individual chunk items have been encoded into WM. In contrast, when a chunk is presented first, the representation of the chunk is loaded into WM, while the individual representations of its elements can be removed in one sweep (“wiping” WM, Ecker et al., 2014). After that, the only representation in WM is the chunk representation. Moreover, the individual representations initially encoded did not damage any other information already stored in WM. In contrast, when the chunk is presented last, the chunk is only retrieved from LTM after all nine letters of the three lists have been encoded into WM. At this point in time, any damage done to previously encoded items is difficult to repair because the items of the chunk have to

be removed individually. Selective removal of individual items is slower than removal of all contents of WM (Ecker et al., 2014) and arguably more error-prone.

Although the benefit of chunks is also assumed to be larger when presented first, a decay-and-rehearsal theory nevertheless predicts that chunks presented last should help memory for earlier presented lists. This is because a single chunk representation requires less time to be rehearsed than the representations of three individual consonants. Hence, even though a chunk should help more for not-chunked information when presented first, it is also assumed to help when presented last. The results of Experiment 2 do not support this claim.

A singleton in the upper row improved memory for subsequent lists more than a chunk in the upper row. This finding is challenging for the assumption in the embedded-process theory that the amount of WM capacity a chunk requires in WM is independent of chunk size (Cowan, 2001).

A potential critique regarding the smaller benefit for chunks than singletons could be that we used known acronyms as chunks. It may be that not all participants were familiar with all acronyms. Therefore, the dependency of the size of the chunk benefit on chunk size may be because participants did not perfectly know the acronyms. However, the fact that the acronyms tended to be recalled even better than singletons suggests that this was likely not the case.

A further critique is that the leading blanks in the Singleton Last condition before presentation of the singleton allowed participants to strengthen memory for the previously presented lists. The same is not possible for chunks because there was no free time in-between presentation of the middle and lower list. In addition, the blanks in between the middle list and the singleton may have made the lists temporally more distinct (Brown, Neath, & Chater, 2007). To control for these possibilities, we replicated Experiment 2 ($n = 20$) but changed the leading blanks to ending blanks following presentation of the singleton in the Singleton Last

condition. This removes any break in-between presentation of the middle and the lower row, thereby removing any effect of temporal distinctiveness and prolonged encoding of the middle-row list before encoding of the lower-row singleton. This modification did not change any of the main findings reported in Experiment 2, namely that singletons but not chunks improved memory for new lists encoded before. The results of this experiment are available in the online supplementary materials.

In the control experiment we observed the same tendency as in Experiment 2 for chunks to be remembered better than singletons. When analyzing the data of both experiments together, there was strong evidence (BF = 15) that chunks were remembered better than singletons. This is a further challenging finding for the embedded-process theory because memory for chunks should be independent of chunk size. At least two interpretations of this finding seem viable. First, chunks are semantically more distinct in a trial than singletons. Representations of the former may have richer associations in LTM because chunks were known acronyms. Hence, visual or semantic representations in LTM may assist recall in addition to the chunk representation in WM. Singletons are certainly less distinct in that sense because they were also the individual elements of the other lists. Second, representations of the individual items forming the chunk may linger in WM due to incomplete removal. In combination with the chunk representation, they may aid recall of the chunk, while at the same time diminishing the chunking benefit for recall of new lists.

A final critique pertains to the type of representations used in Experiment 2 and the control experiment. Chen and Cowan (2009) distinguish between *central* and *phonological* storage, and only central storage is assumed to be limited by a fixed number of chunks. AS is meant to prevent maintenance of phonological representations. Chen and Cowan observed recall of a constant number of chunks only with AS but not without AS. These authors argued that any influence from phonological storage obscures a capacity limit in terms of chunks. It

could be argued that participants in our first two experiments remembered a chunk not only by the chunk representation but also by the phonological representations of the individual items. If that was the case, remembering the new lists may have been more difficult in the chunk condition than in the singleton condition, for example, due to competition for rehearsal, or due to domain-specific interference between phonological representations (see Thalmann & Oberauer, 2017). This difference should disappear if participants only used central storage to remember the chunks and singletons.

5.5. Experiment 3

In Experiment 3, we added AS to restrict the use of phonological representations. We tested whether the benefit of singletons was still larger than the benefit of chunks when phonological storage is prevented. The logic was the following: It is possible that in Experiment 2 singletons helped more than chunks because of the partial reliance on phonological representations, which are more complex, and take longer to rehearse, for three-letter acronyms than for single letters. If that is the case, the differential benefit of singletons vs. chunks should disappear under AS in Experiment 3. However, if we still observe the singleton benefit to be larger than the chunk benefit, we can be confident that the fact that chunks size matters is independent of phonological length and complexity.

5.5.1. Methods

Participants, Materials, and Procedure. Twenty university students (16 women; M ≈ 24 years old) participated in Experiment 3 for one session lasting approximately one hour. Participation was compensated with 15 Swiss Francs or partial course credit. The materials and the procedure were exactly the same as in Experiment 2, except that participants engaged in AS. Participants started to articulate continuously “ba bi bu” at a self-chosen rate before the stimuli were presented and stopped to do so when the first recall cue appeared on the screen.

5.5.2. Results

Serial Recall of Chunked Lists. Memory performance in the five experimental conditions is plotted against item input position in Figure 8 and memory performance for the three list types is plotted against row of presentation in Figure 9. We tested as a manipulation check whether chunks and singletons were recalled more accurately than new lists. On average, recall accuracy was .71, .64, and .47 for singletons, chunks, and for new lists, respectively. Again, the evidence was compelling that singletons were remembered better than new lists (BF = 4.1e+11), and that chunks were remembered better than new lists (BF = 320’185). In contrast to Experiment 2 and the control experiment, chunks were not remembered better than singletons (BF = 0.16). Comparing the average recall accuracies for the three list types with the two previous experiments shows that adding AS decreased memory especially for chunks and new lists, but hardly for singletons.

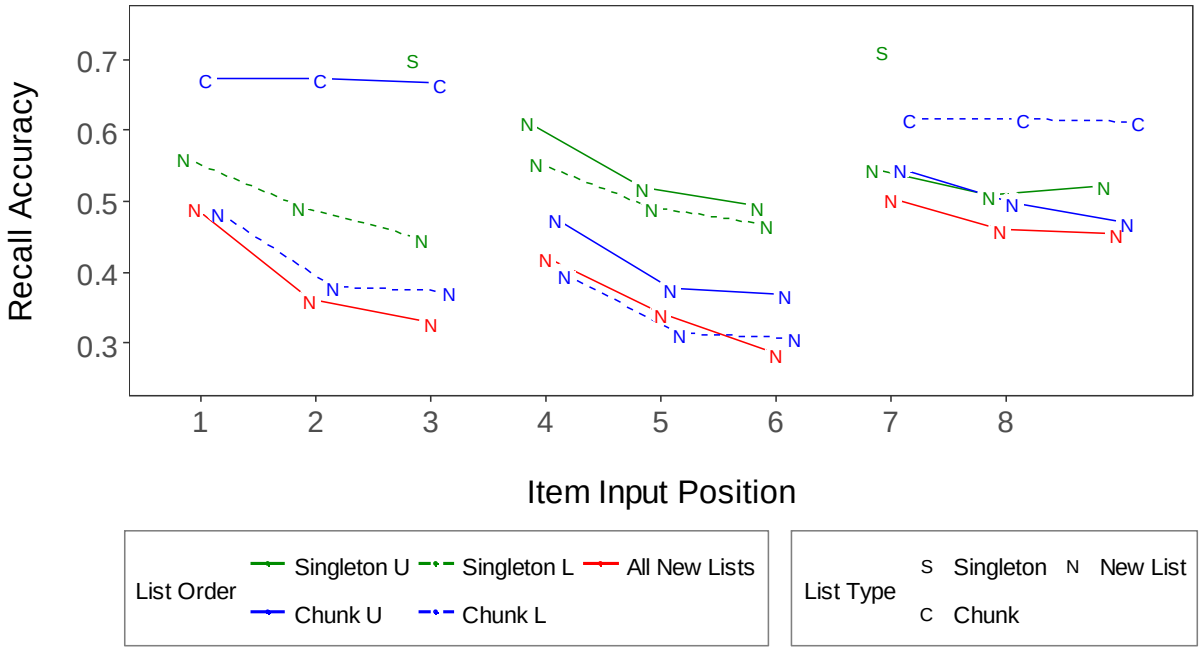

Figure 8. Serial recall accuracy (proportion correct) is plotted against item input position separately for each of the five experimental conditions in Experiment 3. For better interpretability of the figure, (a) we allocated singletons presented in the upper and lower row in the figure to item input position 3 and 7, respectively, and (b) we omitted error bars.

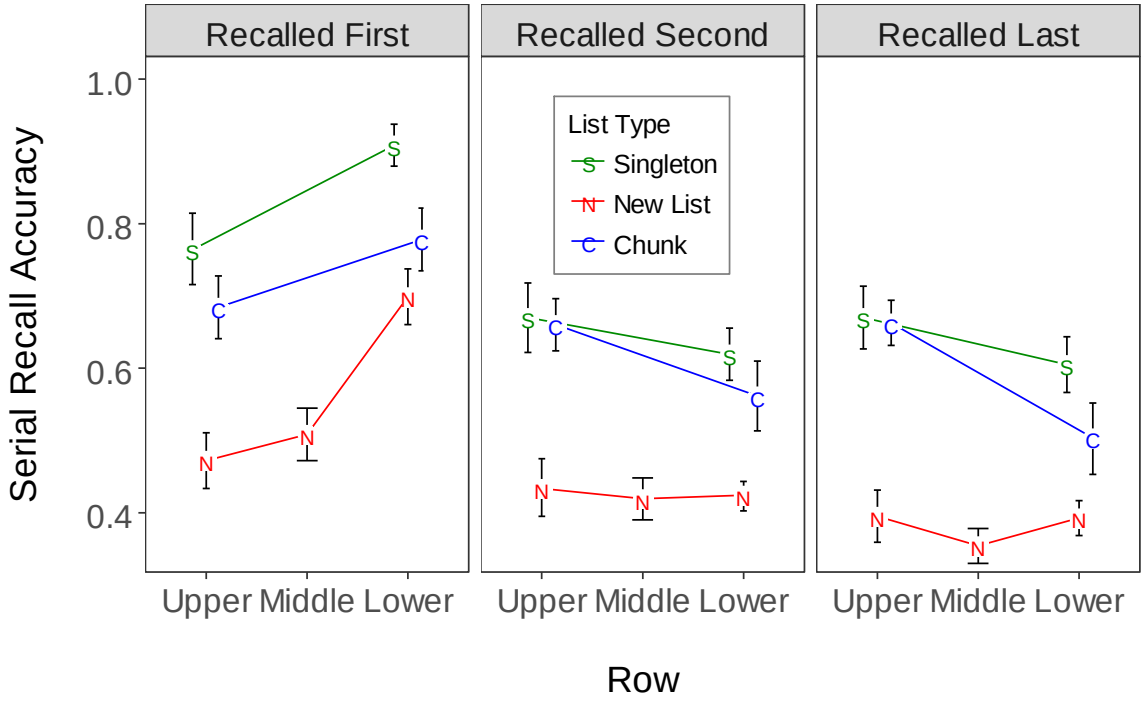

Figure 9. Serial recall accuracy (proportion correct) in Experiment 3 for the three list types plotted against row of presentation. Error bars represent within-subjects standard errors.

Serial Recall of New Lists: Chunk First. We tested with the same Bayesian regression model as in Experiment 2 whether new lists were recalled more accurately than in the control condition when a chunk or a singleton was presented in the upper row. In Figure 10 it appears that both chunks and singletons improved memory for later presented new lists. The analyses (see Table 4) showed, however, that only singletons but not chunks improved memory for later presented new lists credibly. Most importantly for the current purpose, the benefit on later presented new lists was credibly larger for singletons than for chunks. There was no evidence that the effect of having a singleton or chunk in the upper row differed between the middle and lower row (Singleton x Row, Chunk x Row). Finally, there was decisive evidence for output interference: The number of previously recalled lists had a deteriorating effect on new list recall. There was again no compelling evidence that output interference from recalling a singleton or a chunk damaged memory for subsequently tested lists less than output interference from recalling a new list. This suggests that output interference happens at the level of lists and not at the level of individual singletons or chunks.

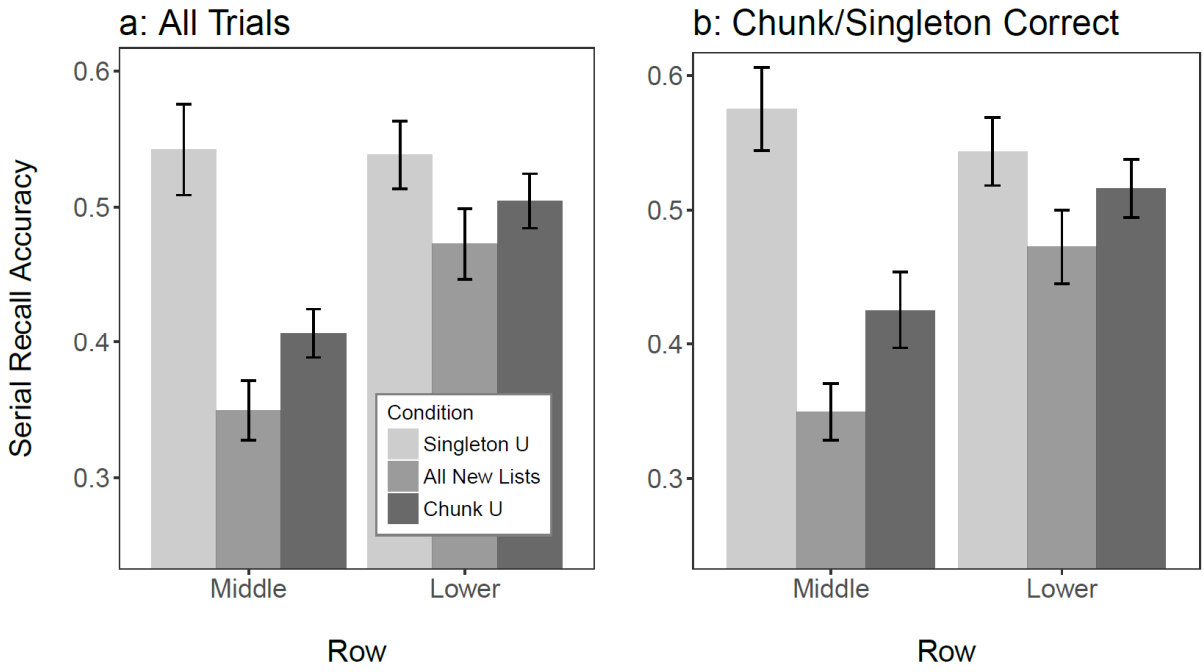

Figure 10. Mean serial recall accuracy of new lists that were preceded by a singleton, a new list, or a chunk plotted against row of presentation in Experiment 3. Panel a is based on all data, and Panel b is based only on those data from the Chunk first condition and the Singleton first condition in which participants recalled the chunk or the singleton correctly. The error bars represent within-subjects standard errors.

Table 4

Posteriors means, 95% HDIs, and the BFs of the parameters of the linear regression model fitted to the data of Experiment 3 in panels a and b of Figure 10, respectively.

		Measure			
		Posterior Mean	95% HDI		Bayes Factor
Chunk First	**Effect**		Lower Bound	Upper Bound	
All Trials	Chunk (vs. Baseline)	0.04	0.01	0.08	0.61
	Singleton (vs. Baseline)	0.13	0.09	0.16	2.20E+09
	Row	0.07	0.00	0.15	0.38
	Chunk x Row	-0.02	-0.09	0.04	0.07
	Singleton x Row	-0.12	-0.24	-0.01	0.85
	Singleton vs. Chunk	0.08	0.05	0.12	116.00
	Nr. Previous Recall Attempts	-0.12	-0.15	-0.09	4.50E+11
	Recalled Chunk Before	-0.05	-0.13	0.03	0.15
	Recalled Singleton Before	0.08	-0.01	0.16	0.34
Chunks and Singletons Correct	Chunk (vs. Baseline)	0.06	0.02	0.12	2.20
	Singleton (vs. Baseline)	0.14	0.10	0.19	7.30E+07
	Row	0.06	0.03	0.14	15.00
	Chunk x Row	-0.02	-0.10	0.07	0.08
	Singleton x Row	-0.16	-0.24	-0.03	62.00
	Singleton - Chunk	0.08	0.04	0.14	13.00
	Nr. Previous Recall Attempts	-0.11	-0.13	-0.08	1.50E+22
	Recalled Chunk Before	-0.02	-0.13	0.09	0.09
	Recalled Singleton Before	0.09	-0.01	0.20	0.42

Serial Recall of New Lists: Chunk Last. Next, we tested whether having a singleton or a chunk in the lower row benefitted recall accuracy of earlier presented new lists. Figure 11 shows better recall on average for new lists compared to the control condition when a singleton followed, but not when a chunk followed. The statistical analysis (see Table 5) confirmed that impression by showing decisive evidence for the comparison Singleton vs. Baseline, but strong evidence for the Null for the comparison Chunk vs. Baseline. The effect of having a singleton or chunk presented in the lower row did not differ between the upper and the middle row (Singleton x Row, Chunk x Row) and there was compelling evidence that having a singleton in the lower row was more beneficial than having a chunk in the lower row. The results regarding output interference corroborate the previous results: The number of

previously recalled lists decreased memory, but type of list did not matter (Recalled Chunk Before and Recalled Singleton Before).

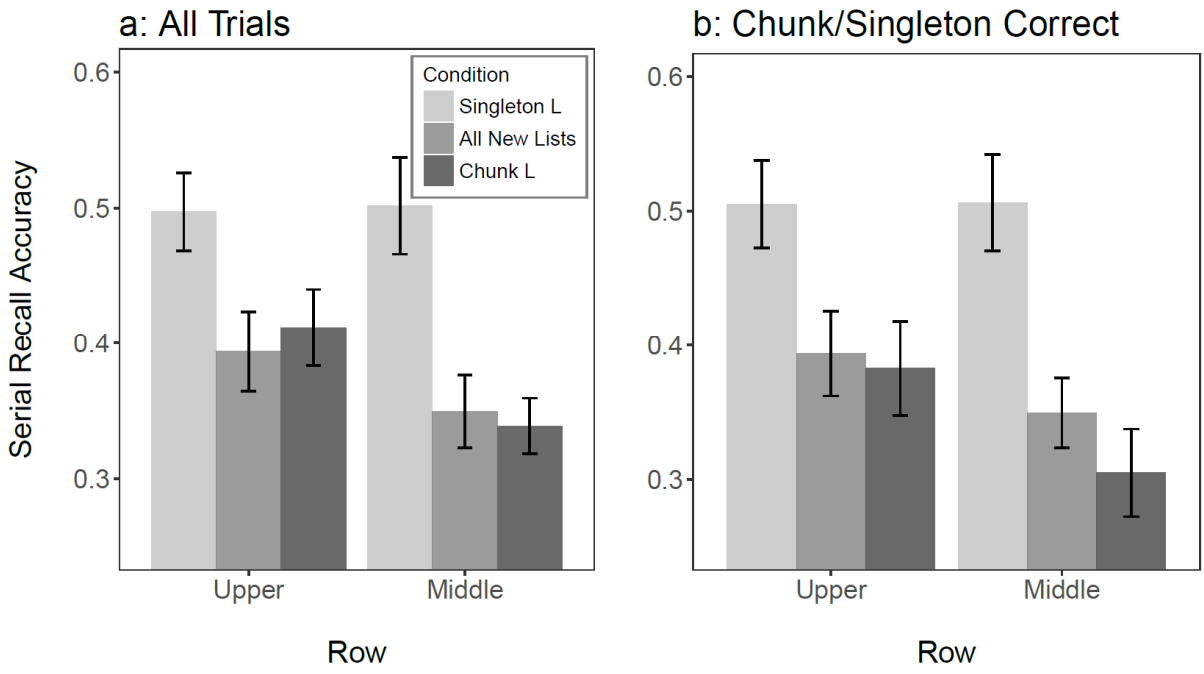

Figure 11. Mean serial recall accuracy of new lists when a singleton, a new list, or a chunk was presented in the lower row (Panel a: full data; panel b: Data from trials with correct recall of chunks/singletons) in Experiment 3. Error bars represent within-subjects standard errors.

Table 5

Posteriors means, 95% HDIs, and the BFs of the parameters of the linear regression model fitted to the data of Experiment 3 in panels a and b of Figure 11, respectively.

		Measure			
		Posterior Mean	95% HDI		Bayes Factor
Chunk Last	**Effect**		Lower Bound	Upper Bound	
All Trials	Chunk (vs. Baseline)	0.03	-0.01	0.06	0.07
	Singleton (vs. Baseline)	0.16	0.12	0.19	1.30E+15
	Row	0.03	-0.09	0.14	0.10
	Chunk x Row	-0.01	-0.07	0.04	0.05
	Singleton x Row	-0.03	-0.09	0.03	0.08
	Singleton vs. Chunk	0.13	0.09	0.18	1’098’477.00
	Nr. Previous Recall Attempts	-0.08	-0.10	-0.06	1.60E+08
	Recalled Chunk Before	0.00	-0.11	0.10	0.09
	Recalled Singleton Before	-0.03	-0.12	0.06	0.09
Chunks and Singletons Correct	Chunk (vs. Baseline)	0.03	-0.02	0.08	0.09
	Singleton (vs. Baseline)	0.17	0.13	0.21	3.80E+11
	Row	0.01	-0.10	0.13	0.10
	Chunk x Row	-0.03	-0.11	0.04	0.09
	Singleton x Row	-0.04	-0.11	0.03	0.12
	Singleton - Chunk	0.14	0.08	0.19	1.00E+04
	Nr. Previous Recall Attempts	-0.08	-0.10	-0.06	2.40E+10
	Recalled Chunk Before	0.01	-0.12	0.14	0.10
	Recalled Singleton Before	-0.04	-0.16	0.07	0.12

5.5.3. Discussion

The main question of Experiment 3 was whether having a singleton in the beginning of a trial still benefitted memory for new lists more than a chunk when participants were required to perform AS. According to Chen and Cowan (2009) only central storage is limited by a fixed number of chunks, and AS forces participants to rely predominantly on central storage. However, even with AS we still observed that the singleton benefit was larger than the chunk benefit. This result further supports our conclusion that the size of a chunk is important in determining its beneficial effect when chunks cannot be maintained by remembering their first element. Likely, the chunk-size effect cannot be attributed to output interference because first recalling a chunk or a singleton decreased memory for about the

same amount as recalling a new list. This finding confirms the prediction by Farrell (2012) according to which output interference happens at the level of lists (i.e., clusters). The main difference to Experiment 2 and the control experiment was that adding AS reduced memory for chunks and new lists, but hardly for singletons. At the moment, we can only speculate why this happened. For example, it could be that participants prioritized retention of singletons over retention of chunks and new lists.

5.6. Experiment 4

Experiments 2 and 3 showed that having a chunk in a trial helped retention of new lists only when the chunk was presented in the upper row, but not in the lower row. We assume that a chunk reduces the load in a WM task only after it has been recognized as a chunk. Only then can participants replace the representations of the individual items with a chunk representation. Clearly, recognizing a chunk in the lower row is not more difficult than in any other row, which is shown by the superior recall of chunks presented in this row compared to new lists in Experiments 2 and 3 (see Figures 4 and 5 and Figures 8 and 9). Apparently the reduction of load afforded by the chunk cannot repair the damage that has been added to representations of previously encoded lists. However, given that the previous experiments only assessed the effects of chunks presented at the beginning or at the end of the memory set (i.e., upper and lower row), we still do not know whether a chunking benefit is observed if a chunk appears mid-way through the trial.

Therefore, the main question addressed in Experiment 4 was how the beneficial effect of chunks depends on their serial position. To attain this goal, we focused exclusively on the chunk conditions (dropping the singleton conditions), and allowed the chunk to appear in all three rows. The previous experiments render two hypotheses plausible. The first possibility is that chunks only yield a benefit for lists encoded after the chunk has been presented, but not

for previously encoded lists. This hypothesis rests on the assumption that chunking can free capacity for encoding subsequent items, but cannot undo the damage to already encoded items. The second possibility is that chunking also helps preceding lists as long as the damage done to them is only mild (i.e., only moderate interference, or reduction of resources, or decay), so that these lists can still be repaired after the load on WM has been reduced. The novel condition, in which a chunk was presented in the middle row, allowed us to test these two hypotheses.

5.6.1. Methods

Participants. Thirty-two university students (22 women; $M \approx 23$ years old) participated in one 1-hour session in exchange for 15 Swiss Francs or partial course credit.

Materials. We used the same set of 30 chunks as in Experiments 2 and 3. Because we dropped the two singleton conditions we only required three times as many new lists as chunks even though chunks could appear now in all three rows. We created a pool of 90 new lists by shuffling the letters of the chunks three times. The algorithm checked that no consonant was used twice within the same list. We changed one or the other letter between new lists manually because chunks were sometimes re-created via shuffling. We again compared chunks and new lists on overall familiarity measured as Google hits (restricted to Switzerland) and on bigram frequency. The Bayesian t-test on $\log_{10}$ Google hits showed decisive evidence (BF = 1.34×10^{31}) that chunks ($M = 5.62$) were more familiar than new lists ($M = 4.02$). Another Bayesian t-test on $\log_{10}$ bigram frequencies showed substantial evidence for the null hypothesis (BF = 6.16) that chunks ($M = 4.21$) did not differ from new lists ($M = 4.20$). All lists were used three times in Experiment 3 and they are available on the OSF webpage.

Procedure. In total, there were four conditions. In all four conditions, three 3-consonant lists were presented sequentially on the screen from top to bottom as in Experiment 2. In the Baseline condition, three new lists were presented for encoding. In the remaining conditions, a chunk was presented either in the upper, middle, or lower row. Every list item was presented for 1 s. There was no time between presentations of two consonants within or between lists. Participants were required to recall the three lists immediately after presentation of the ninth consonant. In all four conditions, the lists in all rows were probed ten times to be recalled first, second, and third. Recall and scoring were the same as previously.

5.6.2. Results

Serial Recall of Chunked Lists. First, we compared serial recall accuracy for chunks and new lists (shown in Figures 12 and 13) with a Bayesian linear regression as a manipulation check. This time, we used the data from all three rows because chunks and new lists could be presented in any row. The analysis indicated that chunks (M = .81) were remembered better than new lists (M = .62), which was supported with a BF = 1.5×10^{24}.

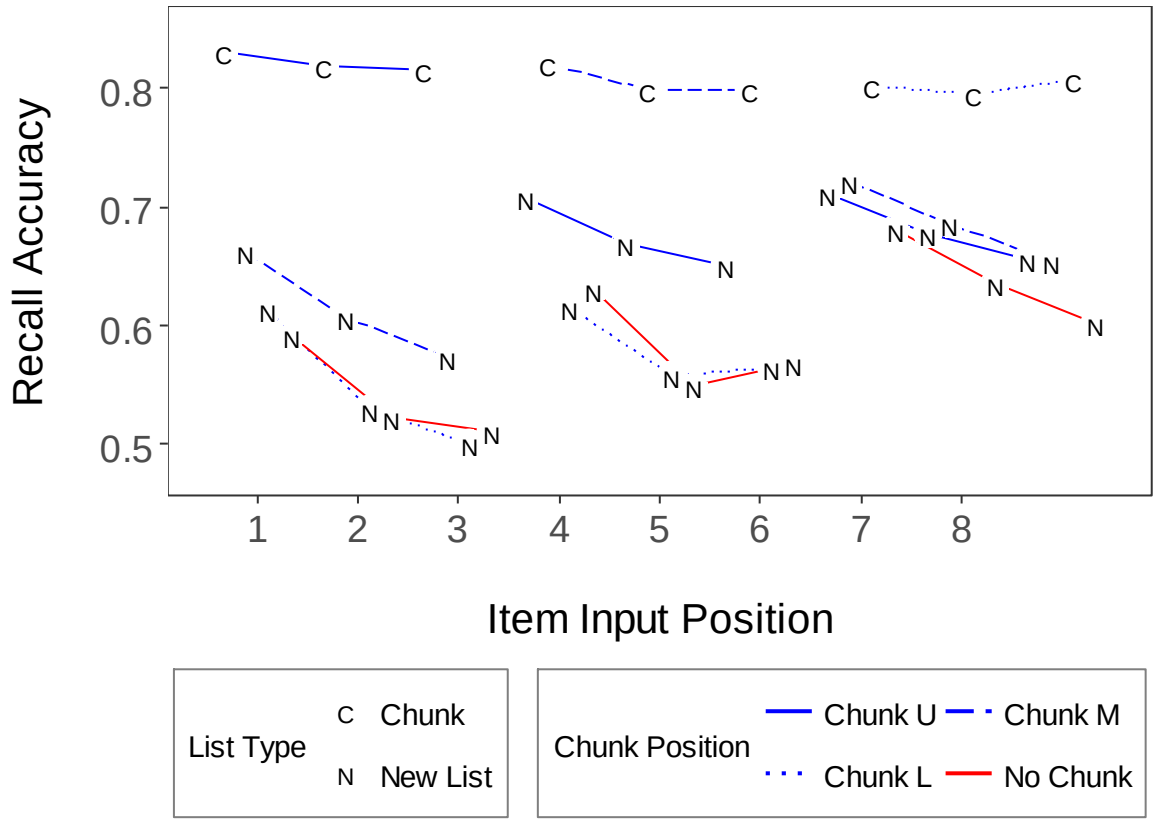

Figure 12. Serial recall accuracy (proportion correct) is plotted against item input position separately for each of the four experimental conditions in Experiment 4. For better interpretability of the figure we omitted error bars.

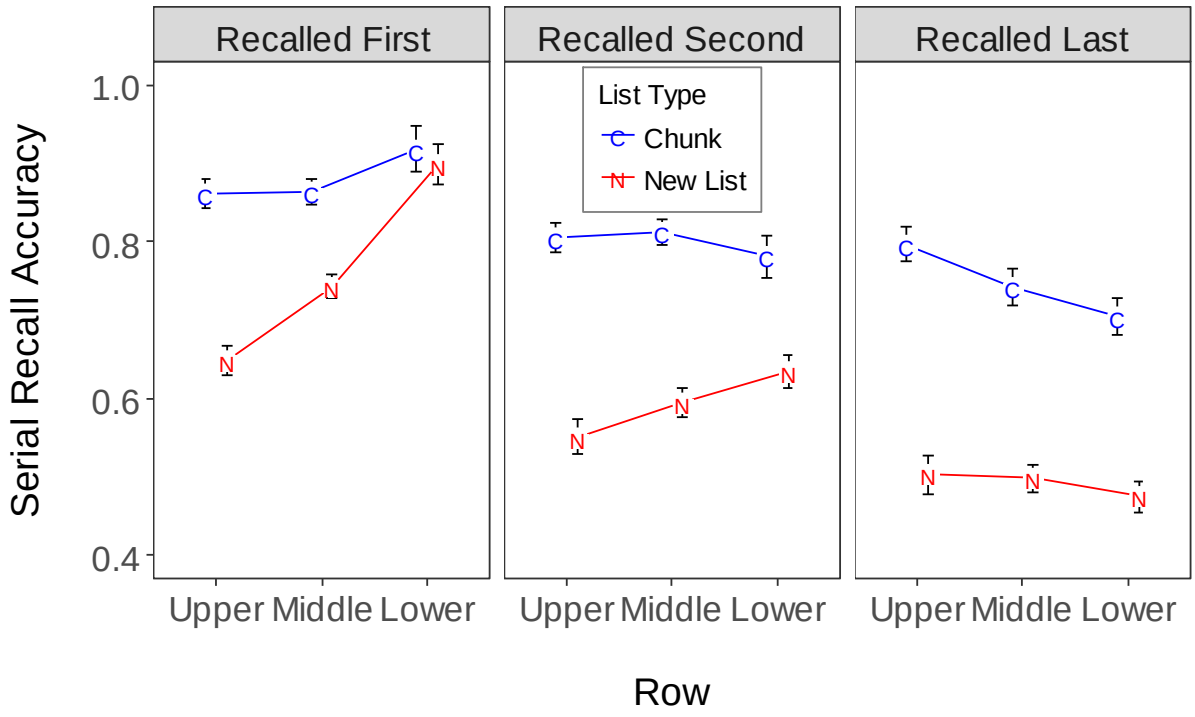

Figure 13. Serial recall accuracy in Experiment 4 for chunks and new lists plotted against row of presentation. The three panels represent the three list output positions.

Serial Recall of New Lists: Chunk Before. Next, we analyzed recall of new lists as a function of whether or not a chunk appeared in a preceding row. Figure 14a shows the data of all trials, and Figure 14b shows data conditioned on the correct recall of the chunk. Serial recall accuracy was analyzed in a Bayesian linear regression using as independent variables row of presentation (middle vs. lower, zero-centered) and condition (Baseline, chunk in the upper row, and chunk in the middle row). The latter variable was entered into the regression as two simple-coded variables using the Baseline condition as the reference category. The first contrast, Chunk in Upper Row, compared the Baseline condition to lists that were preceded by a chunk in the upper row; the second contrast, Chunk in Middle row, compared the Baseline condition to lists that were preceded by a chunk in the middle row (which could only happen for lists in the lower row). In Figure 14 it is visible that a new list was recalled more accurately when it was preceded by a chunk in the upper or middle row. There was

substantial to strong evidence (see Table 6) that a chunk in the upper row increased memory more for lists in the middle row than the lower row.

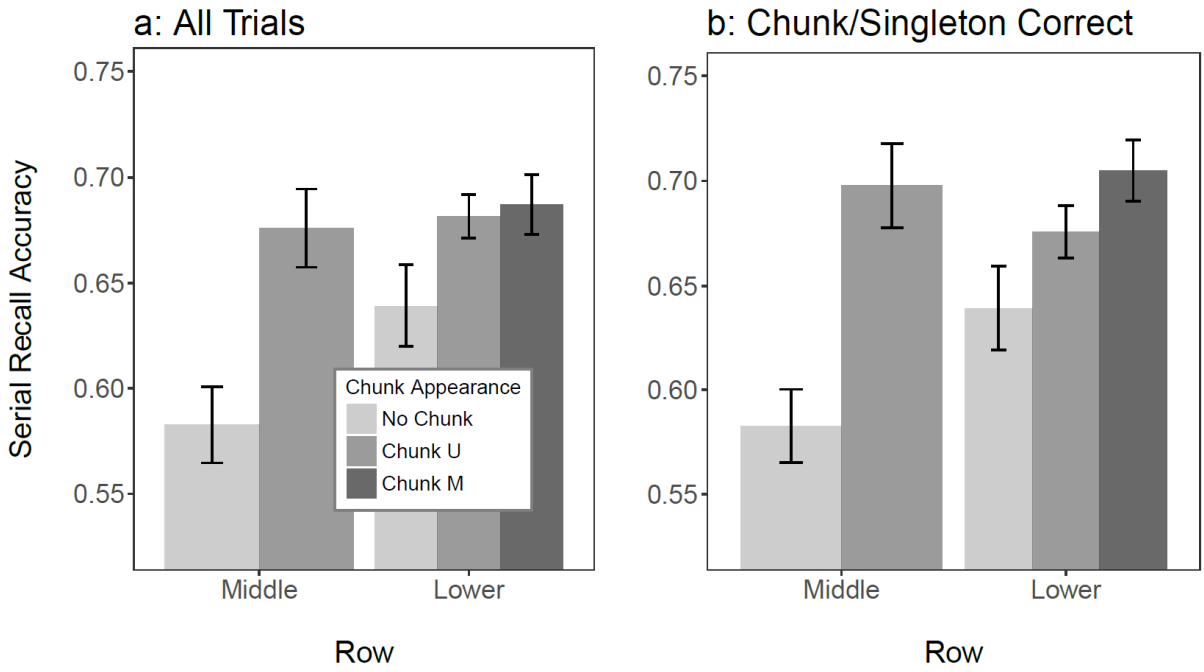

Figure 14. Serial recall accuracy of new lists that were preceded by a chunk compared to new lists in the Baseline condition of Experiment 4. Panel a shows data of all trials, whereas panel b subsets the data of trials with chunks in which chunks were recalled correctly.

Table 6

Posterior means, 95% HDIs, and the BFs of the parameters of the linear regression fitted to the data shown in panels a and b of Figure 14, respectively.

Chunk Before	Effect	Measure			
		Posterior Mean	95% HDI		Bayes Factor
			Lower	Upper	
All Trials	Chunk in Upper Row	0.07	0.04	0.10	825.00
	Chunk in Middle Row	0.05	0.01	0.08	1.20
	Row	0.04	-0.01	0.09	0.12
	Row x Chunk in Upper Row	-0.05	-0.10	0.00	0.37
	Nr. Previous Recall Attempts	-0.17	-0.20	-0.15	3.60E+42
	Recalled Chunk Before	-0.05	-0.10	0.00	0.24
Chunks Correct	Chunk in Upper-Row	0.08	0.04	0.11	878.00
	Chunk in Middle Row	0.07	0.03	0.11	9.20
	Row	0.03	-0.02	0.09	0.09
	Row x Chunk in Upper Row	-0.07	-0.13	-0.02	2.00
	Nr. Previous Recall Attempts	-0.18	-0.20	-0.15	7.90E+43
	Recalled Chunk Before	-0.06	-0.12	0.00	0.41

Serial Recall of New Lists: Chunk After. We analyzed recall accuracy of new lists that were followed by a chunk and compared it with trials in which only new lists were presented (see data on Figures XXa and XXb) using a Bayesian linear regression. Row of presentation (again zero-centered) and condition were used as independent variables. The latter variable was entered as two simple-coded variables. The first variable, Chunk in Middle Row, compared the Baseline condition to the condition with a chunk in the middle row. The second variable, Chunk in Lower Row, compared the Baseline condition to the condition with a chunk in the lower row. The only constellation in which a chunk helped memory for a previously presented list was when the chunk appeared in the middle row: in this case, memory for the list in the upper row improved compared to the Baseline condition. The BFs and the HDIs, which are shown in Table 7, indicate that the beneficial effect of a chunk in the middle row was credible. The evidence was strongly against a beneficial effect of a chunk in the lower row, replicating the result from Experiment 2. There was also evidence against the two-way interaction. Together, the two analyses of memory for new lists depending on chunk

position were in line with Experiments 2 and 3. They supported the claim that the detection of chunks helps remembering new lists as long as WM has not been heavily loaded.

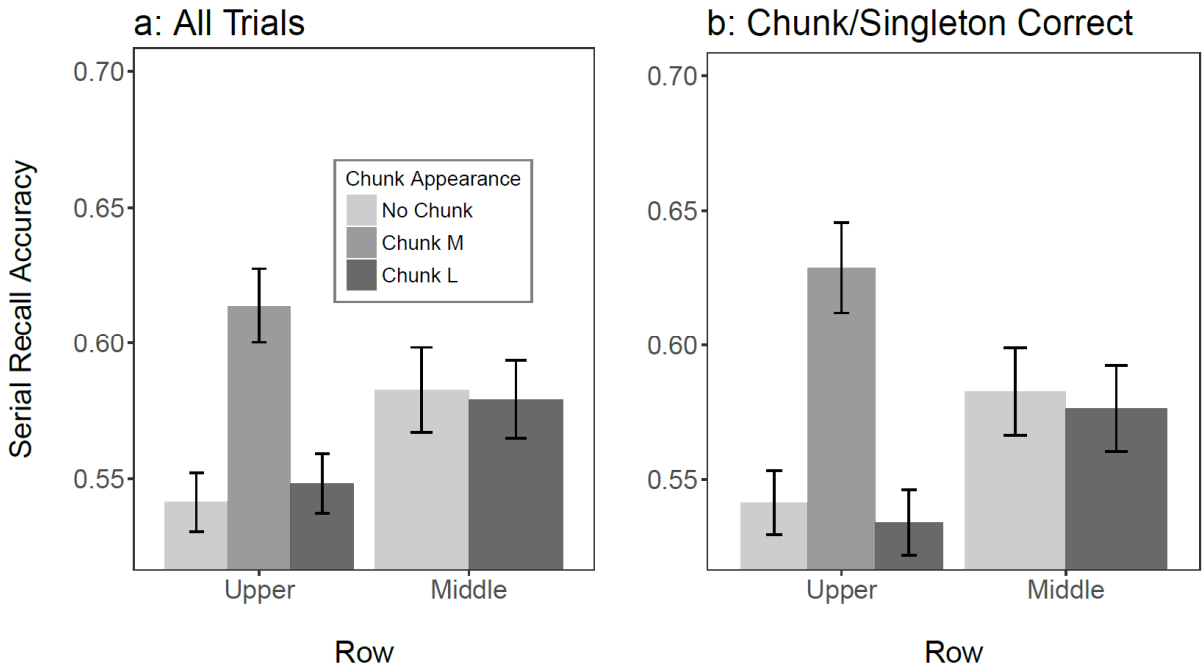

Figure 15. Serial recall accuracy of new lists that were followed by a chunk in Experiment 4. Panel a shows data of all trials, whereas panel b shows data of trials with chunks conditioned on correct recall of the chunk.

Table 7

Posterior means, 95% HDIs, and the BFs of the parameters of the linear regression fitted to the data shown in panels a and b of Figure 15, respectively.

Chunk After	Effect	Measure			
		Posterior Mean	95% HDI		Bayes Factor
			Lower	Upper	
All Trials	Chunk in Middle Row	0.07	0.04	0.11	125.00
	Chunk in Lower Row	0.00	-0.02	0.03	0.02
	Row	0.04	0.01	0.06	1.30
	Row x Chunk in Position 3	-0.01	-0.06	0.04	0.04
	Nr. Previous Recall	-0.09	-0.11	-0.08	4.60E+38
	Recalled Chunk Before	0.01	-0.04	0.05	0.04
Chunks Correct	Chunk in Middle Row	0.09	0.05	0.13	993.00
	Chunk in Lower Row	-0.01	-0.04	0.02	0.03
	Row	0.04	0.01	0.07	1.30
	Row x Chunk in Lower	0.01	-0.05	0.06	0.05
	Nr. Previous Recall	-0.09	-0.11	-0.07	2.50E+25
	Recalled Chunk Before	0.02	-0.03	0.08	0.06

5.6.3. Discussion

Experiment 4 focused on the question at what list positions chunks helped remembering new lists in a trial. By varying the serial input position of the chunk in the trial we evaluated whether a chunk only helps memory for lists that follow it (because it frees capacity to encode those lists) or whether a chunk can help lists that preceded it (because it frees capacity to re-establish degraded representations in WM).

As in the previous experiments, chunks improved memory for the following lists. This was also the case for chunks presented in an intermediate serial position in the memory set (i.e., in the middle row). Presenting a chunk in the middle row also improved memory for the preceding list. After presentation of the middle–row list, WM is already loaded with six consonants. At that point, the representations of the consonants in the first list (upper row) have been degraded from encoding the consonants in the second list (middle row). Nevertheless, recognizing the chunk in the second list helped memory for the first list. If we

assume that representations of different items interfere with each other, the representations of the first list will be distorted by the subsequent encoding of the items of the second list. These distorted representation have to be disambiguated to be recalled, a process which is called redintegration (Lewandowsky, 1999). If the second list items can be replaced by a chunk, the first list items can still be retrieved and redintegrated before they are further distorted by the third list. In addition, if the individual items of the chunked list can be replaced by a chunk representation, the total amount of interference in WM is reduced. Hence, better memory for the first list could be explained if we think of participants redintegrating successfully the first list after recognizing the chunk, thereby reconstructing the representations of the first list's items. Redintegration will be less successful in the Chunk Last condition because more interference between representations has happened after encoding nine items.

5.7. General Discussion

The goal of the present series of experiments was to examine the relation between WM and LTM in serial recall tasks using sequential presentation of stimuli. We evaluated how WM can make use of information stored in LTM by recoding several stimuli into a familiar unit, a process known as chunking.

Evidence that Chunking frees Capacity. Previous work showed that the amount of information recalled in WM tasks could be well described by a fixed number of chunks (Chen & Cowan, 2005, 2009; Cowan et al., 2004). However, these studies left open whether chunks help because they free capacity in WM or because LTM assist reconstruction of chunks at recall (Hulme et al., 1991, 1995). To test whether chunking frees capacity for other, not-chunked information, we tested the chunking benefit on not-chunked information. We consistently observed such a benefit across all experiments, confirming that chunking frees WM capacity.

Chunk Size Matters. If WM capacity is a limit on the number of chunks, chunk size should have no effect on how much capacity a chunk consumes (Chen & Cowan, 2005, 2009). Therefore, our next aim was to test whether the chunking benefit depended on chunk size. It did not in Experiment 1, but it did in Experiments 2, 3, and the control experiment. In the latter experiments, we constructed the chunks so that participants cannot restrict encoding and maintenance to the first element of a chunk. In contrast, the chunks used in Experiment 1 allowed to use this strategy, because the first element was unique to every chunk. With the assumption of that strategy, any theory of WM predicts that chunk size has no influence on the chunking benefit. This follows because participants are not required to encode and maintain the whole chunk, but only a single item. When participants were required to encode all elements of a chunk, as in Experiment 2, there was a smaller chunking benefit for three-element chunks compared to singletons.

The constant capacity in terms of a fixed number of chunks reported in earlier studies (Chen & Cowan, 2005, 2009; Cowan et al., 2004) may to some degree be due to the material that allowed the application of the above mentioned memory strategy. We cannot make this attribution with confidence, however, because the comparison between our Experiment 1 – enabling this strategy – to our Experiments 2, 3, and 4 – not enabling the strategy – is confounded with other differences, in particular pertaining to the materials and the size of the memory sets. Moreover, there are arguably other factors, apart from strategy differences, that contribute to the different results when using chunks with unique or not unique elements. For example, when using chunks with not-unique elements, similarity between two chunks sharing some elements, and between chunks and singletons, is larger than when chunks are composed of unique elements. Similarity between list items increases the chance of confusing them, thereby decreasing serial recall performance(Conrad, 1964; Saito, Logie, Morita, & Law, 2008).

Could similarity-based confusion explain why singletons led to better memory for other items than three-letter chunks did? For such an explanation to work, we would have to assume that not-chunked items are confused more often with three-letter chunks than with singletons. If that were the case, memory for chunks should be worse than memory for singletons. The opposite was the case in Experiment 2 (and the accompanying control experiment).

Could other forms of interference – apart from confusion – explain the differential benefits of singletons and three-letter chunks? If we assume that the representations of three-letter chunks are more complex than those of singletons (i.e., containing more features or components, such as more phonemes or more letters), then they could interfere more strongly with other items by distorting their representations – a mechanism known as interference by superposition (Oberauer et al., 2012).

Any explanation in terms of factors leading to different amounts of interference between representations takes a step away from the notion that WM capacity is limited in terms of a fixed number of chunks. According to such a model – sometimes referred to as "slot model" – performance should be limited only by the number of chunk representations. Variables such as their similarity or complexity should not matter. It is still possible that there is a core capacity limit in terms of chunks, which is obscured by additional mechanisms affecting memory performance. We ruled out one of them, phonological maintenance and rehearsal, in Experiment 3, but there is an infinite number of other auxiliary mechanisms – including interference – that could be added to a discrete-capacity model. The more such additional mechanisms are invoked, however, the more we need to ask whether the additional mechanisms are not sufficient to explain all extant data on their own, without the assumption of a core capacity limit (Navon, 1984).

Chunking is more beneficial for subsequent than for earlier contents of WM. The chunking benefit interacted with serial position of the chunk. Whereas a chunk benefitted other not-chunked information when presented as the first or second list, there was no such benefit when presented as the third list. This is in agreement with the results by Portrat et al. (2016, Figure 8, p. 431). Although Portrat and colleagues did not test the benefit of chunks on not-chunked information formally, their figure suggests that memory for subsequent letters improved when the preceding letters formed a chunk rather than a random set. We assume that chunks only help retention of other items after a compact chunk representation has been retrieved from LTM, and the representations of the individual items have been removed from WM. These processes can only take place after all elements of a chunk have been presented. Up to that point, the representations already in WM will be damaged by the temporary maintenance of the individual elements of the chunk. When the damage to representations of earlier encoded items is only mild, removal of the representations of the individual chunk elements allows reparation of the damage. However, when the damage is severe, reparation is not possible anymore. This is why chunks presented in in the last list position did not lead to a chunking benefit. A decay-and-reactivation theory of WM additionally predicted that chunks help in any list position, because they require less time to be rehearsed in the retention interval than new lists. However, we did not find evidence for this prediction, consistent with evidence that verbal representations do not decay in WM (Oberauer & Lewandowsky, 2013, 2014).

5.8. Conclusion

Chunks reduce the load on WM, thereby improving memory for other information maintained concurrently. The load by a three-letter chunk, however, still exceeds that from a single letter. Hence, a fixed capacity in terms of number of chunks in WM cannot alone explain the chunking benefit. We propose that the chunking benefit results from the following steps: After the individual items have been encoded into WM, potentially interfering with already encoded representations, a matching chunk representation can be detected in LTM.

This representation is retrieved from LTM and encoded into WM, which allows removal of the representations of the individual elements of the chunk. The ensuing reduction of load on WM facilitates subsequent encoding of further information into WM. In contrast, information encoded before the chunk suffers from the initially high load, and the subsequent reduction of load enables only limited repair of that damage.

6. Revisiting the Attentional Costs of Rehearsal in Working-Memory Tasks

Mirko Thalmann, Alessandra Souza, and Klaus Oberauer

University of Zurich

Submission status:

Submitted for publication

Authors‘ contributions:

Mirko Thalmann: Development of the specific research questions, programming the experiments, analyzing the data, writing the manuscript

Alessandra Souza: Co-supervising the project, Discussing the research questions and the results, commenting on the manuscript

Klaus Oberauer: Development of the general research idea, supervising the project, discussing the research questions and results, commenting on the manuscript

Acknowledgement

This research was supported by a grant from the Swiss National Science Foundation to KO (#149193).

6.1.Abstract

There is a recent surge of interest in maintenance processes in working memory, such as articulatory rehearsal, elaboration, and attentional refreshing. Yet, we know little about the central attentional demand of these processes. It has been assumed that articulatory rehearsal does not require central attention at all (Vergauwe, Camos, & Barrouillet, 2014), being in essence a cost-free strategy. In contrast, elaboration and attentional refreshing are assumed to incur large and continuous costs on central attention. We tested these assumptions in three experiments in which participants were presented with a varying number of words to rehearse. Participants were instructed to rehearse the words aloud, or to elaborate them by creating interactive images. Attentional refreshing was examined in a condition in which words were to be maintained during articulatory suppression. During retention participants carried out a series of choice reaction tasks, which were used to measure central attentional demands of the maintenance strategies. Articulatory rehearsal had costs on processing RTs that lasted for 10 s. These costs started to be clearly attributable to central attention when at least four words had to be rehearsed. Elaboration imposed substantial costs on central attention primarily briefly after the memoranda were presented. Finally, maintenance of words during articulatory suppression did not yield persistent costs on central attention, implying that participants did not continuously refresh the words.

All experimental scripts and data sets reported here are available online (https://osf.io/2r87g/?view_only=b7c0d5f6d0de46ca91af14942ed3b95f).

Keywords: Working Memory, Central Attention, Articulatory Rehearsal, Elaboration, Attentional Refreshing

6.2.Introduction

The ability to keep information in mind for carrying out complex tasks relies on working memory (WM). The capacity of WM to keep representations accessible is severely limited. In an attempt to bypass their severe WM capacity limits, people often engage in maintenance processes, which can be generically subsumed under the term rehearsal. Here, we distinguish between three types of rehearsal that have been proposed in the WM literature: articulatory rehearsal, elaboration, and attentional refreshing. The aim of the present study is to investigate the attentional demands of these three forms of rehearsal.

Three Forms of Rehearsal. Articulatory rehearsal is the overt or covert speaking of verbal material to oneself (e.g., Baddeley, 1996). Everyday observation, studies using overt-rehearsal protocols (Rundus & Atkinson, 1970; Tan & Ward, 2008), and laboratory self-report measures (Bailey, Dunlosky, & Kane, 2011; Dunlosky & Kane, 2007) show that people can engage in articulatory rehearsal, and often do so spontaneously. Accordingly, many models of WM attribute to articulatory rehearsal an important function for the retention of information over the short term (e.g., Baddeley, 1986; Camos, Lagner, & Barrouillet, 2009). Whether articulatory rehearsal really helps short-term retention is a matter of ongoing debate (for a critical review see Lewandowsky & Oberauer, 2015). In contrast, the role of articulatory rehearsal in establishing durable representations over the long term has been found to be minimal at best (Greene, 1987, for a review).

Elaboration consists of enriching to-be-remembered items with already existing representations in long-term memory (LTM; Craik & Tulving, 1975). These representations can be of any type, for example semantic or visual. Elaboration also includes the build-up of new relations between the items to be remembered, such as connecting the words of a list to be remembered in a sequence (Craik & Tulving, 1975; Greene, 1987). Laboratory self-report measures (Bailey et al., 2011; Dunlosky & Kane, 2007) show that in about one third of the trials participants spontaneously engages in elaboration during a WM task using words as

memoranda. Elaboration has a beneficial effect on remembering items over the long term (Craik & Tulving, 1975), but whether it has a beneficial effect over the short term is currently unclear. Correlational evidence in support of a beneficial effect of elaboration comes from studies asking trial-by-trial reports of rehearsal strategies: Recall was better for trials in which participants reported using elaboration (Bailey, Dunlosky, & Kane, 2008).

Lastly, *attentional refreshing* involves briefly thinking of a representation in WM, hence bringing this representation to the focus of attention, and thereby extending and/or augmenting its accessibility (e.g., Raye, Johnson, Mitchell, Greene, & Johnson, 2007). Refreshing is assumed to be a sequential process, in which only one item can be refreshed at any point in time (but see Portrat & Lemaire, 2014), and the time required for refreshing one item has been estimated to around 35-50 ms (Vergauwe et al., 2014; Vergauwe & Cowan, 2014). The role of refreshing for maintenance in WM has been mainly inferred from the observation of better memory performance when participants have some free time in between the memoranda or processing episodes in WM (Barrouillet et al., 2004; Camos et al., 2009) tasks. Only a few studies actually manipulated attentional refreshing as an independent variable in a WM task and examined its effects on short-term memory performance (Souza & Oberauer, 2017; Souza, Rerko, & Oberauer, 2015; Vergauwe & Langerock, 2017). For example, Souza, et al. (2015) investigated refreshing by asking participants to attend to individual WM items during the retention interval. That study showed that refreshing led to better visual WM performance than not refreshing, and this improvement increased linearly with the number of refreshing attempts an item received.

Attentional Costs of Rehearsal. Here, we are interested in the degree to which articulatory rehearsal, elaboration, and refreshing demand central attention. Central attention is a capacity-limited processing mechanism that has been characterized as a bottleneck that enforces serial processing (Pashler, 1994), or a processing resource that constrains the speed of parallel processes (Tombu & Jolicoeur, 2003). Dual-task studies have shown that central

attention is involved in response selection and retrieval from LTM (Johnston, McCann, & Remington, 1995). Here we ask to what extent the three forms of rehearsal rely on central attention.

Knowledge about the central attentional demands of different forms of rehearsal plays an important role in determining how they can be applied in a WM task. Moreover, an understanding of how much central attention is needed for a specific rehearsal strategy is essential in predicting how these strategies can be combined. For example, the time-based resource-sharing (TBRS) model of WM assumes that articulatory rehearsal and refreshing can be applied (Camos, 2015; Camos & Barrouillet, 2014; Camos et al., 2009; Mora & Camos, 2015; Mora & Camos, 2013) simultaneously in WM tasks. The assumption within the TBRS model is that articulatory rehearsal does not require central attention at all, being in essence a cost-free strategy. In contrast, refreshing is assumed to depend on central attention. Therefore, carrying out other attentionally demanding tasks either postpones refreshing, or execution of these tasks is postponed because refreshing is taking place (hence showing a trade-off). If, however, articulatory rehearsal is not attentionally demanding, it can be applied simultaneously with refreshing with no trade-offs.

The assumption that articulatory rehearsal does not demand (or demands very little) central attention rests on the findings of two early studies (i.e., Guttentag, 1984; Naveh-Benjamin & Jonides, 1984). In the following, we will briefly summarize these studies and discuss why conclusions from them regarding the central attentional demands of articulatory rehearsal are problematic.

Guttentag (1984) examined the trade-offs between carrying out overt articulatory rehearsal of a memory list simultaneously with a finger tapping task. Specifically, he compared how many times per minute children from different age groups were able to tap with their forefinger in a single-task condition, and in a dual-task condition simultaneously requiring articulatory rehearsal. A core result of his experiments was that articulatory

rehearsal led to severe costs on the tapping task. The dual-task costs correlated negatively with the age of the children. The interpretation of Guttentag was that articulatory rehearsal becomes more and more automatized with age, which is why his paper is often cited in support of articulatory rehearsal being non-demanding in adults. However, the dual-task cost of articulatory rehearsal in the oldest children (mean age = 11.5 years) was still around 15%, clearly speaking against complete automatization of articulatory rehearsal.

Naveh-Benjamin and Jonides (1984, Experiment 1) used a sophisticated design to examine several questions concerning articulatory rehearsal and elaboration. Participants were given three two-digit numbers on every trial to remember for immediate recall (pre-load task). Next, two words were presented to be rehearsed during retention of the numbers. One group was instructed to rehearse synchronously with a metronome (articulatory rehearsal group); the other group was instructed to elaborate the words independently of the metronome beat (elaboration group). Whereas the elaboration group was informed about a delayed recognition test of the words at the end of the experiment, the articulatory rehearsal group was not. A dot appeared on the screen either 0.850 s, 4.675 s, or 12.325 s after presentation of the two words (hence while participants were rehearsing or elaborating the two words). Participants were instructed to press a button as soon as they detected the dot on the screen. After that, recall of the three two-digit numbers was required, ending the trial. The most important result for the present examination is that the reaction time (RT) to detect the dot was on average 22 ms faster when it appeared 12.325 s after start of rehearsal than when it appeared 4.675 s thereafter. There was no such decrease in RTs over time for the elaboration group. Based on these results, the authors concluded that articulatory rehearsal – but not elaboration – can be executed without the requirement of attention after an initial set-up stage.

There are several reasons why these two studies do not allow clear statements about the central attentional demands of rehearsal to be made. First, both studies assessed dual-task costs on tasks that do not require response selection. Work with the psychological refractory

period (PRP) effect has shown that simple RT tasks such as the ones employed in these studies do not engage the central attentional bottleneck (Pashler, 1994). Based on this research, it is difficult to argue that what was measured by the RT tasks in Guttentag (1984) and Naveh-Benjamin and Jonides (1984) was central attention. Nevertheless, even keeping this caveat in mind, the dual-task cost of rehearsal on tapping in Guttentag's study was substantially larger than zero across all age groups, which seems incongruent with the assumption that articulatory rehearsal is cost-free.

Second, in the experiment by Naveh-Benjamin and Jonides there was no control condition to compare the effects of rehearsal to. Therefore, it is unclear whether dual-task costs remained for articulatory rehearsal even after 12 s or more. Accordingly, this study may have led to an underestimation of the attentional cost of articulatory rehearsal. Third, the requirement to rehearse in synchrony with a metronome could have compromised the measurement of attentional costs of articulatory rehearsal in Naveh-Benjamin and Jonides's (1984) study. The longer RTs to dots presented earlier during the rehearsal time could reflect some initial attentional cost of adapting the pace of articulatory rehearsal to the metronome. Similarly, the fact that RTs in the elaboration group did not decrease over the processing phase may be because participants did not have to align their rehearsal with the metronome. To summarize, the evidence that articulatory rehearsal does not require central attention based on these two studies is weak, at best.

The Present Study. The concerns raised above indicate that it is time to revisit the attentional demands of rehearsal. In the present experiments we made an effort to obtain a proper estimate of the central attentional requirements associated with articulatory rehearsal and elaboration while avoiding the pitfalls previously mentioned. In addition, we assessed the attentional demands of a condition in which participants were required to perform articulatory suppression (AS) during maintenance of a memory list. Vergauwe et al. (2014) have argued that blocking the use of articulatory rehearsal with AS prompts participants to resort to

refreshing of the memoranda. Our goal was to evaluate whether attentional costs in this condition are consistent with the idea that participants spontaneously engage in refreshing.

Experiment 1 examined the central attentional demands of articulatory rehearsal and elaboration in a design closely modeled after Naveh-Benjamin and Jonides (1984). The benefit of this design is that participants executing articulatory rehearsal do not anticipate the memory test for the rehearsed material, and therefore have no incentive to engage in additional processing of the rehearsed material. In this way, central attentional requirements of articulatory rehearsal can be measured while controlling for any additional type of rehearsal that participants may spontaneously use. In Experiment 2 we zoomed in on articulatory rehearsal because of the importance of the assumption that it is a cost-free strategy in the TBRS model (e.g., Camos et al., 2009). We investigated the attentional costs of articulatory rehearsal with a paradigm previously used by proponents of the TBRS theory to investigate attentional costs of maintenance processes (Vergauwe et al., 2014). In Experiment 3, we again compared articulatory rehearsal and elaboration with each other, and we assessed whether people spontaneously engage in refreshing under AS.

In all three experiments we used a choice RT (CRT) task to probe the attentional demand because this task requires response selection, and hence requires central attention (Pashler, 1994). Furthermore, we applied the overt rehearsal methodology (Rundus & Atkinson, 1970) to check for compliance with the articulatory rehearsal instruction. Without requiring overt responses, it is difficult to make inferences about the use of articulatory rehearsal. Moreover, we tested for LTM at the end of the experiments to assess whether participants engaged in elaboration.

To foreshadow our results, articulatory rehearsal yielded dual-task costs on CRT, which increased with the number of items to be rehearsed. The costs of articulatory rehearsal vanished after about 5 s when no WM test was forthcoming. When a WM test was expected, the costs persisted until the end of the processing period (10 s). Elaboration did not have a

long-lasting effect on CRTs irrespectively of WM test expectancy. Similarly, the AS condition supposed to engender refreshing did not yield persistent dual-task costs on CRTs (cf. Vergauwe et al., 2014).

6.3.Experiment 1

In Experiment 1 we gauged the attentional costs of articulatory rehearsal and elaboration in an adapted version of Naveh-Benjamin and Jonides's (1984) experiment. As in their seminal study, we assigned participants to one of two groups: Articulatory Rehearsal or Elaboration. In each trial, participants were presented with 2 words, which they had to read aloud. Thereafter, the Articulatory Rehearsal group was instructed to continuously rehearse the words aloud, whereas the Elaboration group was instructed to create vivid and interactive mental images of the meaning of these words. To control for compliance with the instructions, we recorded the speech of participants in the Articulatory Rehearsal group. Moreover, we presented a delayed recognition test of the memoranda at the end of the experiment. In keeping with Naveh-Benjamin and Jonides (1984), the Elaboration group, but not the Articulatory Rehearsal group, was informed at the beginning of the experiment about the final memory test. This information served as a motivation to carry out the elaboration strategy. Better delayed recognition performance of the Elaboration group in comparison to the Articulatory Rehearsal group is expected if they complied with the instructions.

To assess the time course and magnitude of the attentional costs of articulatory rehearsal and elaboration, offset of the words was followed by a 10-s processing phase in which dots were presented at unpredictable, irregular periods. Our testing of the attentional costs of rehearsal deviates from Naveh-Benjamin and Jonides (1984) in five regards. First, a CRT was used instead of a simple RT (SRT) to measure central attentional demand. Second, a control condition was included in which no words had to be rehearsed at all, hence serving as a single-task baseline for the CRT task within each group. Third, the processing phase in

every trial lasted 10 s and was divided into five segments. Within each segment, a CRT stimulus was shown with a fixed probability. Fourth, we omitted the pre-load of three two-digit numbers used by Naveh-Benjamin and Jonides (1984) to avoid any possible confounds between the costs of rehearsing the words and any other potential rehearsal participants would attempt to maintain the pre-load memoranda. Fifth, we also dropped the requirement for participants to articulate in synchrony with a metronome. The requirement to align a speech output with a metronome may need central attention, which would have compromised the attempt to measure the central attentional demands of articulatory rehearsal alone.

To summarize, Experiment 1 had three independent variables: rehearsal instruction (articulatory rehearsal vs. elaboration, varied between-participants), set size (two vs. zero words to be rehearsed, varied within participants), and time segment within which a processing stimulus was presented (1-5, varied within-participants).

6.3.1. Method.

Participants. Fifty-four university students (34 women; *M* = 26 years old) were randomly assigned to one of two groups: Articulatory Rehearsal or Elaboration. In this and the remaining experiments in this article, sample size was determined a priori to values that in our experience are sufficient to obtain robust evidence for medium-sized effects in within-subjects (and mixed) designs. All participants in the experiments reported here were compensated with partial course credit or 15 Swiss Francs for participation in one 1-hour session. All participants were university students and native speakers of German. They provided written informed consent and were debriefed in the end of the experiment. Moreover, participants were informed that their speech during the experiments would be recorded and inspected to control for compliance with the instructions.

Materials and Procedure. All experiments reported here were programmed in MATLAB using the Psychophysics Toolbox 3 (Brainard, 1997; Pelli, 1997). Participants

were tested in individual booths where they sat at a distance of approximately 50 cm from the computer screen. They wore headphones equipped with a microphone for recording of their speech.

For each participant, 100 two-word lists were constructed by randomly selecting (without replacement) from a pool of 200 mono- and disyllabic German nouns. Forty control lists with the letter string “xxxxx” were added to this pool, resulting in 140 lists. On every trial of the rehearsal task, one list was randomly sampled without replacement from this pool.

In every trial, the two rehearsal words were presented simultaneously in the center of the screen for 600 ms followed by a blank delay of 400 ms (see Figure 1a). Next, a 10-s interval followed in which the rehearsal task was combined with a CRT processing task (described below). Participants in the Articulatory Rehearsal group were instructed to uninterruptedly rehearse the words aloud during the processing phase. Participants in the Elaboration group were instructed to continuously elaborate the words during the processing phase for a subsequent delayed memory test. Participants were informed that in the 2-word condition their main task was to rehearse the two words according to their group's instruction, and their secondary task was to respond to the dot task. In the 0-word condition, when the “xxxxx”-string was presented on the screen, they were instructed to simply respond to the dot task (0-word condition); in the Articulatory Rehearsal group they were instructed to remain silent during that time. The 0-word condition served as a single-task baseline of responses to the dot task.

As in Naveh-Benjamin and Jonides’s (1984) study, the Elaboration group was informed that elaboration is particularly helpful for remembering the words in the long-term. More specifically, participants in this group were told to: *“Create a mental image of the two words. Then, try to make an image incorporating both images, such that they interact with each other. And try continuously to change their interaction or make the image more vivid.”*

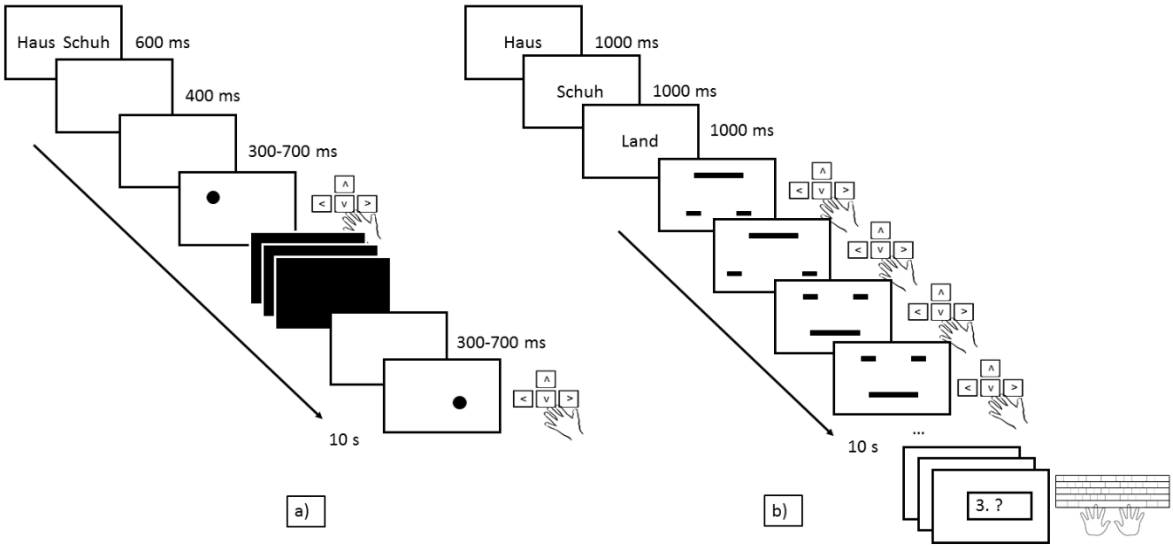

Figure 1. Illustration of the sequence of events in a trial in Experiment 1 (panel a) and Experiments 2 and 3 (panel b). Panel a: In the beginning of each trial, two words for the rehearsal task were displayed simultaneously. Next, a 10-s period started in which participants were instructed to continuously rehearse the words while concurrently performing a dot CRT task. The dot task was divided into five 2-s segments. In each segment, a dot was presented with p = 0.46. The black frames represent segments of 2000 ms, in which no processing stimulus happened to be presented. Panel b: In Experiments 2 and 3, the memory words were presented sequentially. The first stimulus of the CRT was presented immediately after offset of the last word, and each CRT response was immediately followed by the next CRT stimulus until 10 s had elapsed. After that, participants were requested to attempt typed forward serial recall of the words. Stimuli are not drawn to scale.

In the processing task, participants had to indicate as fast and as accurately as possible whether a dot (occurring at unpredictable intervals) was shown above or below the horizontal screen midline by pressing the up or down arrow keys on the keyboard. The total duration of the processing task was 10 s, which was divided into five 2s-segments. Within every segment, a dot (diameter = 60 pixels) was shown with p = 0.46. To create uncertainty about the dot's location, its horizontal position was randomly selected from a uniform distribution ranging from -250 to +250 pixels in relation to the center of the screen. The vertical position of the dot was selected such that 4/5 of the dot (48 pixels) fell either in the upper or lower part of the screen. If a dot was selected to be shown within a given 2-s time segment, its onset after the beginning of the segment was sampled from a uniform distribution between 300 and 1000 ms. The dot remained visible until an answer was given or after 500 ms of the next time segment

had elapsed. An answer was counted as valid if it occurred while the dot was onscreen; otherwise, a time-out was recorded. In case an answer to a presented dot was not given in the same time segment but in the 500 ms of the next time segment, and another dot was scheduled to be presented in that time segment, the overlap of that response into the new time segment was added to the onset time of the next dot (which was again sampled from a uniform distribution between 300 and 1000 ms). The overlap was however not added to the total duration of the processing task, which did never outlast 10 s. This procedure assured that time pressure was within a reasonable range.

After completion of the 140 trials of the main experimental task, a delayed recognition test of the rehearsal words was presented for all participants. The main purpose of this test was to ensure that the Elaboration group complied with the instruction to elaborate on the presented words. For the Articulatory Rehearsal group, the delayed recognition test was a surprise. The first five and the last five rehearsal word lists were excluded from this test to control for primacy and recency effects. One noun of each of the remaining 90 lists was randomly selected as a cue. Each trial in the delayed recognition test consisted of the presentation of a cue on the left side of the screen, and a set of four candidate words (the correct word and three distractors) on the right side of the screen. One distractor was a word presented in another memory list (intrusion), and the remaining two distractors were new words selected from a pool of 180 mono- and disyllabic German nouns. Participants were instructed to select among the set of candidate words the one that was presented together with the cue during the rehearsal task. Participants responded by clicking with the left mouse button on one of the four words. The order of testing of the rehearsal words was randomly determined for each participant.

6.3.2. Results

Statistical Framework. We used Bayesian statistics for all analyses. Bayesian statistics overcome some of the shortcomings associated with conventional null-hypothesis significance testing (e.g., Wagenmakers, 2007). In the Bayesian framework (for an introduction see Kruschke, 2014), prior knowledge about the credible values of the model parameters is expressed as probability distributions known as priors. These priors are updated in light of the data to yield posterior distributions. The posterior reflects the knowledge about credible values of model parameters after taking into account the data.

In addition to the posteriors of credible model parameters, the relative credibility of two models can be computed with the Bayes Factor (BF). The BF is the ratio of the marginal likelihoods of two models (e.g., Rouder, Morey, Speckman, & Province, 2012). By formulating two hypotheses of interest as statistical models and assuming equal priors for the two models, the BF quantifies how many times more likely one hypothesis is than the other, given the data. For the comparison of a null hypothesis with an alternative hypothesis on a parameter of interest (e.g., the difference between two groups in a t-test, or an effect in an ANOVA design), the point null hypothesis is represented by a model in which the prior of the relevant parameter is set to 0. The interpretation of the BF is straightforward. A BF of 1 states that the data are ambiguous, providing no evidence favoring one model or the other. Although a BF between 1 and 3.2 states that one model is more likely than the other, it is usually “not worth more than a bare mention” (Kass & Raftery, 1995). A BF from 3.2 to 10 is regarded as substantial evidence, a BF from 10 to 100 as strong evidence, and a BF larger than 100 as decisive evidence for one model over the other.

Compliance with task instructions. Six participants were excluded due to responding close to chance level in the CRT task. They fell below the 99.9% quantile of a binomial distribution with mean 0.5 (i.e., less than 190 correct responses across the 325 CRT episodes presented over the experiment). Another participant in the Articulatory Rehearsal group was

excluded because this participant mentioned after the experiment to have expected the delayed recognition test and used elaboration to better remember the words. Six additional participants were excluded, because they did not adhere to the rehearsal instructions, as assessed by the inspection of the speech recordings of the trials. For example, some participants articulated the “xxxxx”-string in the control condition instead of remaining silent, some read wrong words or only sporadically rehearsed some of the word lists. This resulted in a final sample of 41 participants, with $n = 23$ in the Elaboration group and $n = 18$ in the Articulatory Rehearsal group.

Delayed Recognition. Next, we analyzed the responses in the delayed recognition test to assess whether the elaboration instruction worked. If participants in the Elaboration group complied with the instruction, we expect to observe better LTM for the rehearsal word pairs compared to the Articulatory Rehearsal group, replicating the results of Naveh-Benjamin and Jonides (1984). Responses in the delayed recognition test were classified into three categories: hits (selection of the correct alternative), intrusions (selection of the alternative with a word from another list), and false alarms (selection of one of the new words). The proportion of responses in each of these categories is shown in Table 1. Hits (log-odd transformed) in the Articulatory Rehearsal condition and the Elaboration condition were compared with a Bayesian linear regression model.

Table 1
Proportion of Responses in each Response Category in the Delayed Recognition Test in Experiment 1

	Condition	
Response Category	Articulatory Rehearsal	Elaboration
Hits	0.425	0.818
Intrusions	0.293	0.095
False Alarms	0.141	0.043

Note. False-Alarm rate was divided by 2, because the probability of making a False Alarm was twice that of making a Hit or an Intrusion.

The regression model was run via JAGS (Plummer, 2003), which was accessed via the R statistical computing environment (R Core Team, 2017) with the *rjags* package (Plummer, Stukalov, & Denwood, 2015). We approximated the BF of the difference between conditions via the Savage-Dickey density ratio. This method obtains the BF for two nested models, such as a Null model assuming that the effect is zero, and an alternative model that allows the effect to vary freely. The first step is to obtain the posterior of the parameter of interest – here, the size of the difference between conditions in the probability of a hit – in the alternative model. The BF is then obtained by dividing the height of the posterior by the height of the prior at the parameter value assumed in the Null model (for further details see Lee & Wagenmakers, 2014; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009). The BF was 8.42×10^6 for the Alternative hypothesis over the Null, which is decisive evidence for better delayed recognition in the Elaboration group than in the Articulatory Rehearsal group.

Processing Task. RT and accuracy to respond in the processing task served as the dependent variables. RTs were trimmed as follows. First, RTs associated with incorrect answers and time-outs were removed (15.57% of all RTs available for analysis). Second, RTs that exceeded or fell below the individual mean ± 3 standard deviations in each time segment

were excluded (0.87% of the remaining RTs). The remaining RTs were averaged within each segment in each set-size condition (2 words vs. 0 words) and group (Articulatory Rehearsal vs. Elaboration). In Figure 2, RTs (panel a) and accuracies (panel b) are plotted against segment. RTs and accuracies (log-odd transformed) were analyzed separately with 2 x 2 x 5 Bayesian ANOVAs using the BayesFactor package 0.9.10-2 (Morey & Rouder, 2014), which is available in R. For all analyses with the BayesFactor package in this article we used the default combination of JZS priors (Cauchy prior on effect size, Jeffreys prior on the variance).

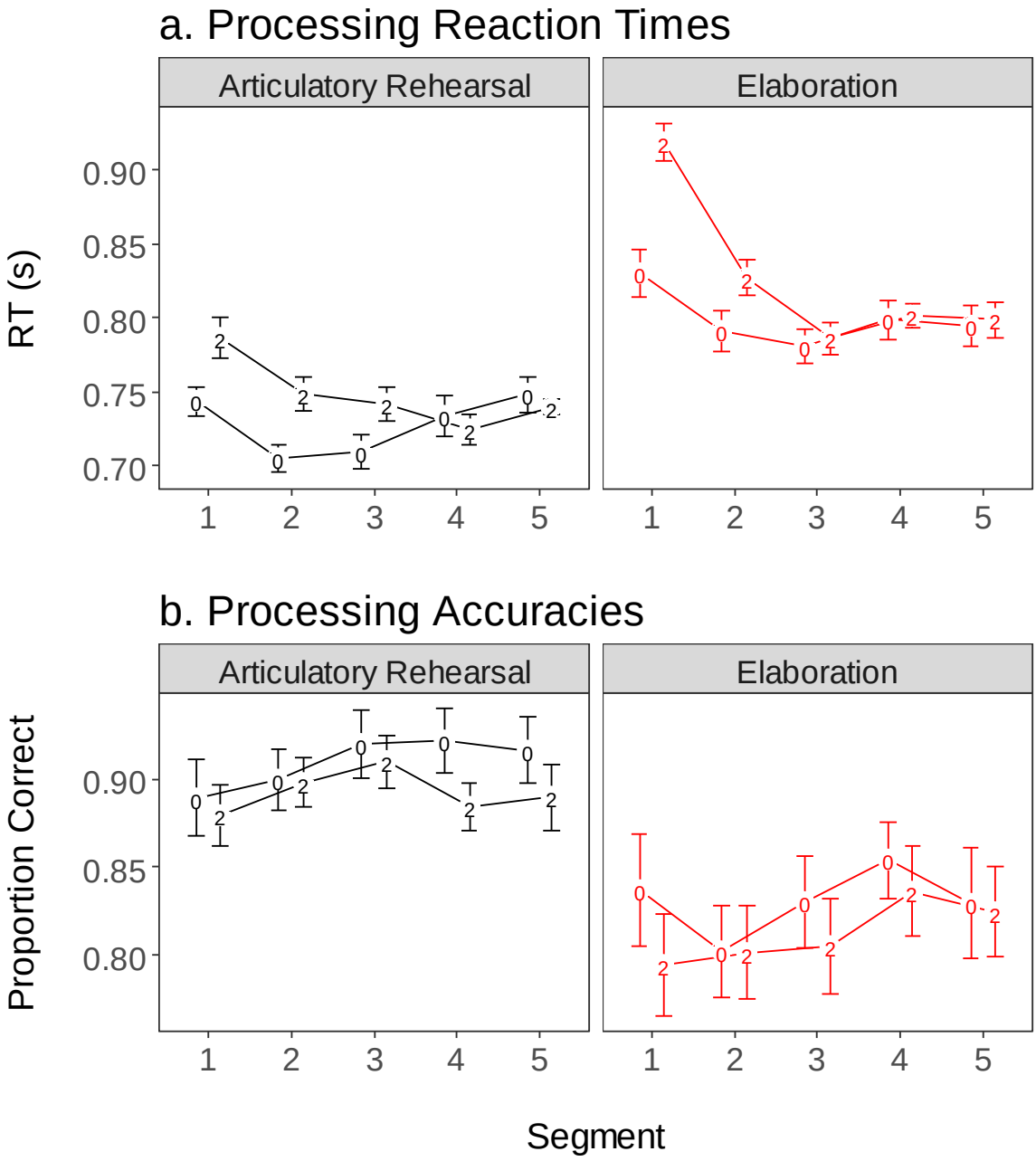

Figure 2. RTs (panel a) and accuracies (panel b) plotted over segment for each set size and rehearsal group in Experiment 1. Error bars represent standard errors for within-subjects designs.

For each ANOVA, the winning model was selected as the one with the highest BF in comparison to the null model. Evidence for each effect included in the winning model was gauged by comparing the winning model to a model derived from it by dropping the effect in question. Conversely, evidence against an effect excluded from the winning model was assessed by adding it to the winning model.

For RTs, the winning model included fixed effects of segment, group, set size, set size x segment, and a by-subject random intercept (BF = 6.11×10^{13} over the null model). Next, we describe the evidence for each effect of theoretical relevance. The Elaboration group responded more slowly than the Articulatory Rehearsal group (BF = 3.8). Rehearsal slowed responding in the processing task, as indicated by the slower RTs in the 2-words condition compared to the 0-word control condition (BF = 458). Moreover, RTs decreased across the five time segments (BF = 1.39×10^{9}). The costs of rehearsal (2-words vs. 0-words) decreased over segment (segment x set size, BF = 71.6). There was not enough evidence to support a difference between the Elaboration group and the Articulatory Rehearsal group on the decrease of RTs over segment (segment x group, BF = 1.7). The evidence was against the two-way interaction of group x set size (BF = 0.19; hence indicating that the Null was supported by a factor of 5.3). This indicates that the time cost resulting from rehearsing 2 words versus not rehearsing at all did not differ between the two rehearsal instructions. Moreover, there was also evidence against the three-way interaction (BF = 0.16, evidence for the Null equals 1/0.16 = 6.25), indicating that the decrease in the costs of rehearsal over segment was similar for both rehearsal groups.

For accuracies, the winning model included the fixed effects of rehearsal group and of set size, and a by-subject random intercept (BF = 11’238 over the Null). The BF in favor of rehearsal instruction was 4.7, indicating that participants in the Articulatory Rehearsal group responded more correctly, on average, than participants in the Elaboration group. The higher accuracies and faster RTs in the Articulatory Rehearsal group compared to the Elaboration group rule out a speed-accuracy trade-off to explain the group difference. The main effect of set-size yielded a BF = 2397, which implies that rehearsing 2 words credibly reduced accuracy in the processing task. The BF was 23 against the set size x segment interaction, which is strong evidence that the set-size costs on accuracy did not decrease over the processing period.

6.3.3 Discussion

Two observations support the conclusion that participants complied with the rehearsal instructions. First, elaboration led to far better performance in the delayed recognition test than articulatory rehearsal. Second, inspection of the recorded speech confirmed that the participants in the Articulatory Rehearsal group were articulating the words continuously.

The main findings of Experiment 1 can be summarized as follows. First, responses in the processing task were initially slower when participants had to rehearse 2 words compared to the control condition with 0 words. This shows that articulatory rehearsal and elaboration require central attention for some period of time. Second, the attentional costs of both forms of rehearsal decreased over time. Visual inspection suggests that the time costs of both types of rehearsal vanished after about five seconds; the costs on accuracy, however, were unaffected by time. Third, the magnitude of the attentional costs of rehearsal did not differ between the two rehearsal groups, showing that both articulatory rehearsal and elaboration engage central attention to a similar extent.

Using an optimized experimental design to measure central attentional demands, our results confirm the most important conclusion of Naveh-Benjamin and Jonides (1984) that articulatory rehearsal is initially attentionally demanding, but the demand declines within the first 5 s, after which the RT costs disappear. Costs on accuracy, however, remained. Different from Naveh-Benjamin and Jonides (1984), the present results suggest that the attentional cost of elaboration in terms of RTs likewise declines over the first 5 s elaboration. Taken together, our findings suggest that both examined rehearsal strategies always incur an attentional cost. This cost is unlikely to vanish during a typical WM task such as complex span (in which memoranda and processing episodes are interleaved), because a new rehearsal process has to be initiated every time a new item is presented, and that usually happens at a rate of 1 to 5 s.

We observed an unexpected difference in RTs and accuracies between the two groups: The Elaboration group was slower and less accurate than the Articulatory Rehearsal group.

This difference cannot be attributed to rehearsal of items presented on the current trial because it was also observed for the control conditions (i.e., no interaction of rehearsal instruction x set size). One potential cause of this group difference is that participants in the elaboration group, expecting the final recognition test, engaged in additional elaboration of items presented on previous trials. This could also explain why RT costs persisted only in the elaboration group but not in the articulatory rehearsal group in Naveh-Benjamin and Jonides (1984). We further examined this issue in Experiment 3.

The observation of central attentional demands of articulatory rehearsal clashes with the conclusions reached in a recent study by Vergauwe et al. (2014). Vergauwe and colleagues presented a varying number of verbal memoranda followed by a 12 s processing period. These authors did not observe a delay of responding to the very first CRT in a trial for memory set-sizes of up to four words, compared to a 0-words control condition. They assumed that the absence of the set-size effect on CRT was due to participants maintaining the words through articulatory rehearsal. However, in their study, participants were not explicitly instructed to cumulatively rehearse the words, and no record of their articulatory rehearsal was made. This makes it impossible to verify whether participants in their experiments were actually engaging in articulatory rehearsal, or any other form of rehearsal. In the present experiment, in contrast, participants were explicitly instructed to rehearse the words, and we did observe a rehearsal cost on processing RTs, especially the first one. Our findings suggest therefore that participants in the Vergauwe et al. (2014) experiments may not have rehearsed the words at all.

One may wonder, however, whether the attentional costs of articulatory rehearsal as estimated in Experiment 1 are specific to the design employed in this study. Experiment 1 did not require immediate recall of the words in the end of the trial, the words were presented simultaneously, and the processing stimuli were presented only sporadically. In addition, participants had to rehearse two words at most, which can be considered as a rather small

memory load. This contrasts with the more typical WM set-up used by Vergauwe et al. (2014), which consisted of a Brown-Peterson WM task with a processing task that had to be carried out continuously during the retention interval. To firmly establish the generality of the attentional costs of articulatory rehearsal, in Experiment 2 we assessed the attentional costs of articulatory rehearsal using the experimental set-up of Vergauwe et al (2014).

6.4.Experiment 2

Experiment 1 showed that articulatory rehearsal is not a cost-free strategy. Given the importance of this cost-free assumption in the TBRS theory, in Experiment 2 we sought to replicate the finding of sustained attentional demands of articulatory rehearsal in a more typical WM task set-up. For that purpose, we combined the overt rehearsal protocol of Experiment 1 with a Brown-Peterson WM task as used by Vergauwe et al. (2014). We chose to model Experiment 2 closely on the one of Vergauwe et al. (2014) because the main conclusion from Experiment 1 contradicts their conclusion that articulatory rehearsal does not require central attention.

The goals of the present experiment were two-fold. First, we aimed at replicating the attentional costs of articulatory rehearsal in a more typical WM set-up. Second, we increased the maximal number of words to be rehearsed and manipulated set size in a more fine-grained manner by asking participants to hold between 0 and 4 items in WM. This allowed us to better estimate the attentional cost of adding each additional word to the rehearsal set.

6.4.1. Methods

General Design. On every trial, a variable number of words had to be retained in WM for a subsequent forward serial order recall test. Thus, different from Experiment 1, participants knew that they had to remember the words over the short term. Set size (from 0-4) and the number of syllables of the words in each memory list (mono- or disyllabic) were

independently varied as within-subjects factors. Presentation of the words was followed by a 10 s processing phase in which participants responded to a CRT task. Participants were instructed to cumulatively rehearse the memoranda aloud throughout the processing phase. In contrast to Experiment 1, CRT stimuli were presented continuously during the processing phase, imposing a demand on central attention with high temporal density.

Participants. Twenty-one university students (17 women; *M* = 24 years old) took part in Experiment 2. One participant was removed from the final analyses because rehearsals were not recorded due to experimenter error and hence we could not check for compliance with the instructions. This resulted in a final sample of 20 participants.

Materials. For each combination of set-size (1-4) and number of syllables (one- or disyllabic), twelve wordlists were constructed, yielding 96 trials. Wordlists were constructed for every participant by randomly selecting words (without replacement) from either a pool of 120 monosyllabic or a pool of 120 disyllabic German words. We added 24 trials with memory load of 0, resulting in a total of 120 trials.

The processing task in Experiment 2 comprised a visuospatial fit judgment task similar to the one employed by Vergauwe et al. (2014). In this task, participants had to judge whether a horizontal bar fitted in the space between two dots (see Figure 1b). The horizontal distance between the two dots was randomly sampled in every processing episode with replacement from a set of six values. These six values were obtained by dividing the width of the screen by six values equally spaced from 15-30 (including 15 and 30). The bar was on each side either 10 pixels longer or shorter than the distance between the two dots. For every decision, it was randomly selected whether the bar appeared above or below the two dots to reduce the probability of the same constellation to appear on several subsequent processing decisions.

Procedure. The sequence of events in each trial is illustrated in Figure 1b. Every trial began with the presentation of a message announcing the number of words to be remembered

on the forthcoming trial. Participants self-initiated the trial by pressing the spacebar. After a 1000 ms blank interval a fixation cross was presented in the middle of the screen for 500 ms, followed immediately by the onset of the first word (for trials with set size > 0). Words were presented sequentially in the center of the screen for 1000 ms. There was no blank interval between the presentations of two successive words. Participants were instructed to articulate the words aloud at the time of their presentation. After presentation of the last word, they were instructed to rehearse all words aloud in cumulative forward order until the end of the processing period. They were asked to remain silent in trials with 0 words.

Directly after presentation of the last word (or after the fixation cross in 0-word trials), the first processing stimulus was shown. Participants were instructed that their main task was to remember the words in their serial order with high accuracy. At the same time, they should try to respond as fast and accurately as possible to the processing task. The processing period lasted exactly 10 s. Each processing stimulus remained onscreen until participants responded by pressing the left or right arrow keys to indicate a fit or not-fit response, respectively. Each response was followed immediately by the presentation of the next processing stimulus. Participants responded to as many processing stimuli as possible within the 10-s period.

At the end of the 10-s processing period, participants were prompted to recall the words in the order of their presentation using the keyboard. We required the participants to type only the first three letters of each word to minimize their need for typing. They were consecutively cued with the position of the next word to be recalled. The typed letters were shown on the screen, and typos could be corrected by using the backspace key. When satisfied with their answer, participants pushed the return key to confirm their response. Upper and lower case was irrelevant for scoring the responses. The experiment started with 5 practice trials with set size ranging from 0 to 2, which were excluded from the final analyses.

6.4.2. Results.

Only RTs in trials in which all words were recalled in their correct serial position at the end of the trial entered the analysis. This criterion resulted in the exclusion of 4.54% of all trials (representing 5.68% of trials with set size > 0). From this pool, RTs associated with incorrect processing responses were excluded (3.44% of remaining RTs). Hence, accuracy in the processing task was generally high. An analysis not reported here showed that accuracy in the processing task did not vary credibly with set size. Next, very long RTs (> 5s) were also removed (0.01% of the RTs). Mean RTs are plotted against processing position in Figure 3.

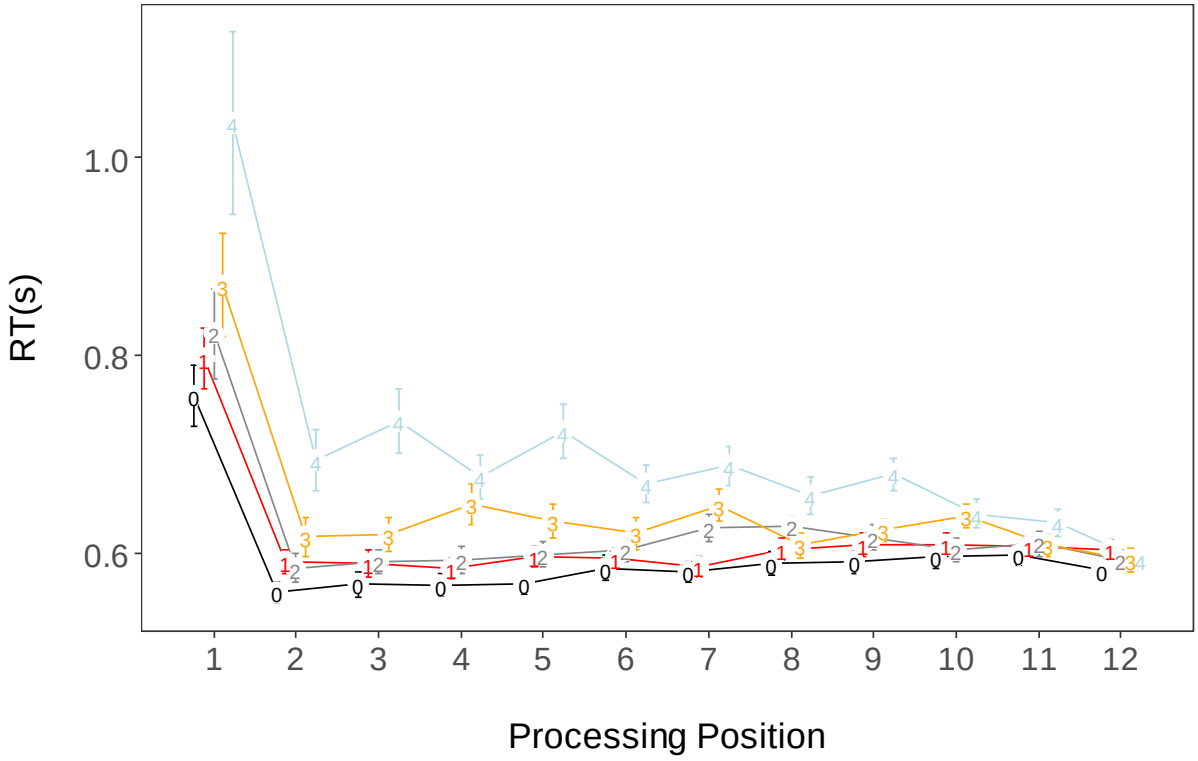

Figure 3. Mean RTs over processing position presented separately for each set size. Processing position counts the answers to the different stimuli presented during the 10-s processing phase. We only plotted data until processing position 12 because after that there were only few observations. Error bars represent standard errors for within-subjects designs.

Three sets of responses were used for assessing the costs of articulatory rehearsal over the processing period: (a) responses to the very first processing stimulus in each trial (henceforth referred to as first RTs), which were analyzed separately because they partially reflect the switch costs between encoding the memoranda and preparing for the processing task; (b) average RTs to all subsequent processing stimuli in the trial (henceforth subsequent RTs), which allow for the estimation of the sustained costs of rehearsal if there is one; and (c) response to the very last processing stimulus in each trial (henceforth referred to as last RTs). We analyzed the last RTs separately to test whether central attentional costs of articulatory rehearsal persist even when a substantial amount of rehearsal has happened. Figures 4a, 4b, and 4c show first, subsequent, and last RTs, respectively, as a function of set size.

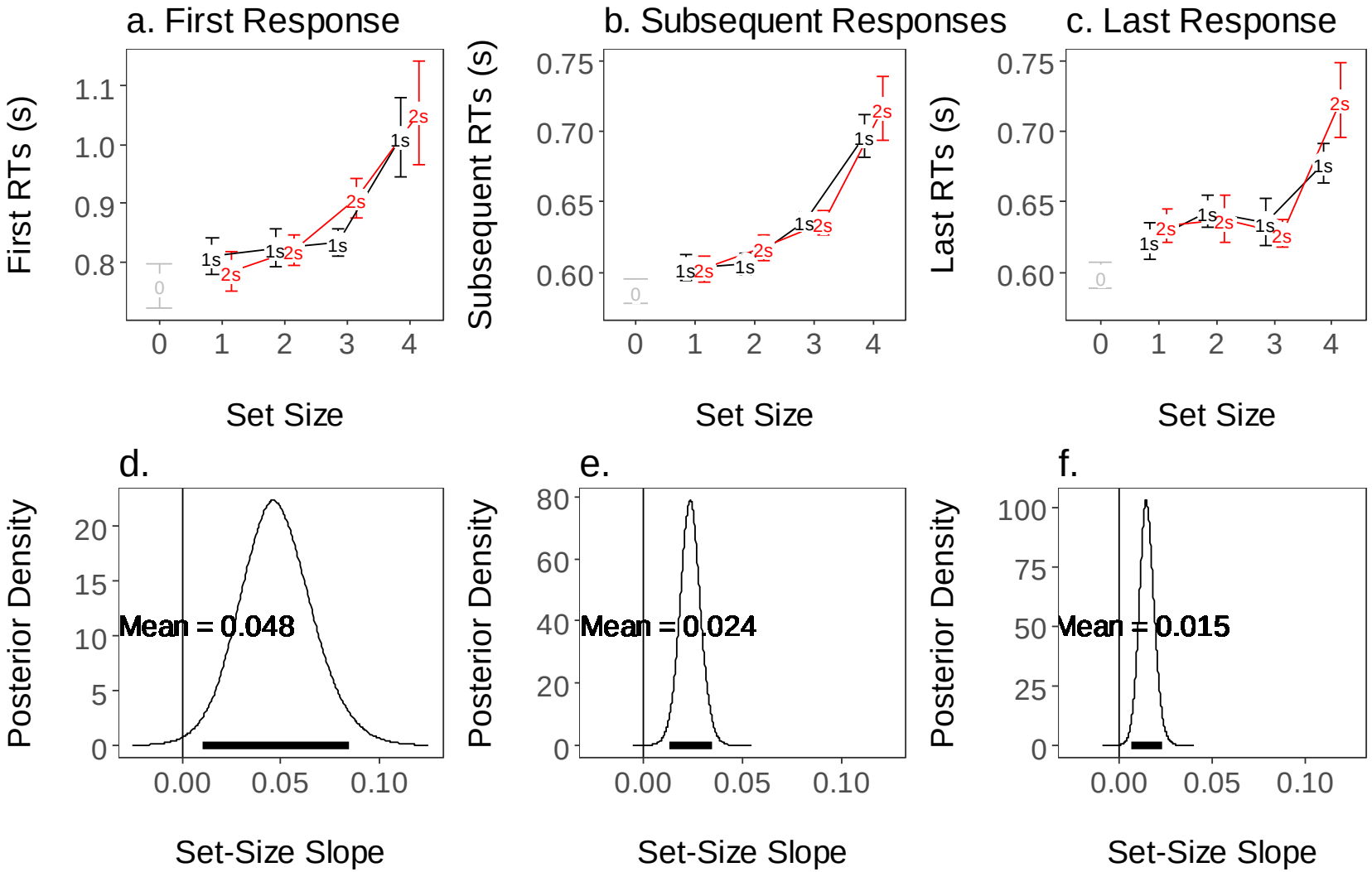

Figure 4. First RTs (panel a), subsequent RTs (panel b) and last RTs (panel c) for mono- and disyllabic words plotted as a function of set size in Experiment 2. Panels d-f show the posterior distributions of the linear set-size effect for first, subsequent, and last RTs, respectively. Error bars in a – c represent standard errors for within-subjects designs. The black bars in the bottom of d – f represent 95% HDIs.

Number of syllables was excluded from the analyses, because an initial analysis (not reported here) revealed substantial evidence against an effect of number of syllables (BF = 7.7, 6.7, and 9.1 in favor of the Null for the analysis of first, subsequent, and last RTs, respectively). This initial analysis already implies that any potential costs of articulatory rehearsal are not related to the duration or complexity of articulation (Baddeley, Thomson, & Buchanan, 1975; Service, 1998). We will come back to this issue in the Discussion.

The regression models were run via JAGS. The graphical representation of the models and the respective priors are shown in Figure 5. The model included a fixed effect of set size as a continuous predictor, a random intercept, and a random slope for the effect of set size, allowing the set-size effect to differ between participants (for the importance of including random slopes see Thalmann, Niklaus, & Oberauer, 2017). We approximated the BFs of the fixed set-size effect via the Savage-Dickey density ratio. For models including random slopes, we additionally report the deviation information criterion (DIC, Spiegelhalter, Best, Carlin, & van der Linde, 2002). Here, we report DICs for models including the set-size effect in comparison to models omitting it (see Table 2). The model comparisons support a linear slowing of processing RTs as a function of set size in all three processing positions. Hence, there was clear evidence for a linear set-size effect on processing RTs for the whole 10-s processing period.

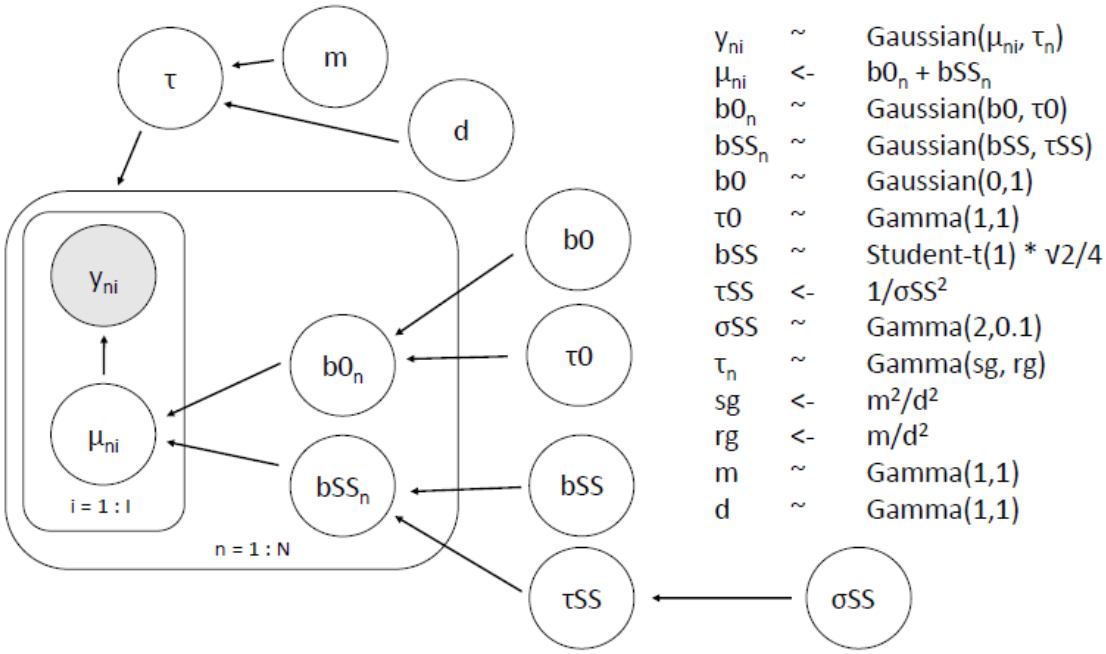

Figure 5. Graphical representation of the regression models run in JAGS used to predict RTs in Experiment 2. The circle shaded in gray represents the data, the circles without shading represent variables to be estimated. The n-plate indicates independent repetitions over N participants, and the i-plate over I trials. SS = set size.

Table 2
BFs and DICs for the models including the set-size effect over the Null model (i.e., omitting it).

	Measure	
Processing Position	BF	ΔDIC
First	3	-112
Subsequent	2330	-247
Last	53	-35

Note. ΔDIC (Difference) = DIC (set size) – DIC (Null). Because smaller DIC values reflect better fit, ΔDIC values below zero favor the set-size model.

Inspection of panels a – c of Figure 4 suggests that the delaying effect on processing RTs is disproportionately driven by set size 4. Up to set size 3 the set-size effect is much flatter. To compare the present results with the contrast of set-size 0 vs. 2 in Experiment 1, we additionally analyzed whether there is a credible difference between set sizes 0 and 2 for first and subsequent RTs, using the same models as above but entering set size as a categorical predictor.

The BFs in favor of the set-size effect were 0.85 and 3.90 for first and subsequent RTs, respectively[1]. The average costs for rehearsing two items were 45 ms and 22 ms for first and subsequent RTs, respectively (i.e., on average 23 ms and 11 ms per additionally rehearsed item). Hence, the additional analyses cast doubt on whether rehearsing only two words is attentionally demanding.

To get a more detailed picture of the central attentional demands of rehearsal we applied the EZ diffusion model (Wagenmakers, Van Der Maas, & Grasman, 2007) to subsequent RTs[2]. The EZ diffusion model computes three psychologically interpretable parameters that allow us to analyze the source of the cost of rehearsal on concurrent processing. The diffusion model assumes that the response selection in a two-CRT is the product of a noisy evidence accumulation process with mean accumulation speed called drift rate (v). Drift rate informs about the quality of the evidence that is accumulated, with a higher drift rate implying better quality. Once the accumulation process hits one of two boundaries (one for the correct, and one for the incorrect response), the corresponding response is given. The distance between the two boundaries is called boundary separation (a) and informs about response caution: A larger boundary separation reflects a more conservative speed-accuracy trade-off setting, resulting in slower but more accurate responses. Finally, the time required for stimulus encoding and executing a motor response is represented in T_{er}. If the slowing of processing as a function of set size is due to a speed-accuracy trade-off, this should be reflected in the a parameter. If not, we should see an effect on T_{er} due to a postponement of the response-selection process, or an effect on v due to the slowing down of the response-selection process.

[1] The BFs in favor of the set-size effect were 1.85 and 383 for first and subsequent RTs, respectively, when comparing set sizes 0 and 3.

[2] The EZ diffusion model requires a substantial number of trials per condition; therefore we could not apply it to the first or last RTs of the present experiment, or the RTs from Experiment 1.

We obtained the BFs for the set-size effects on *v*, *a*, and T_{er} (plotted in Figure 6) by comparing a linear model including that effect (as a categorical predictor with random slopes) to one excluding it. The BF in favor of the set-size effect on *v* and *a* was decisive (9617 and 4134, respectively), whereas the BF for a set-size effect on T_{er} was not strong (5.5). Taken together, the diffusion-model parameters suggest qualitatively different effects of rehearsing small vs. larger word sets on choice RTs. Rehearsal of up to two words at best led to a small postponement of the response-selection process, as expressed in the slight increase of T_{er}. Rehearsal of three or more words, in contrast, resulted in a reduced drift rate (*v*), reflecting impoverished evidence accumulation during response selection. The concomitant increase in the caution parameter (*a*) probably reflects the fact that participants adjusted the boundary separation to the reduced drift rate to ensure a high level of accuracy. At the same time, T_{er} appeared to decline at the higher set sizes; we have no plausible explanation for that effect.

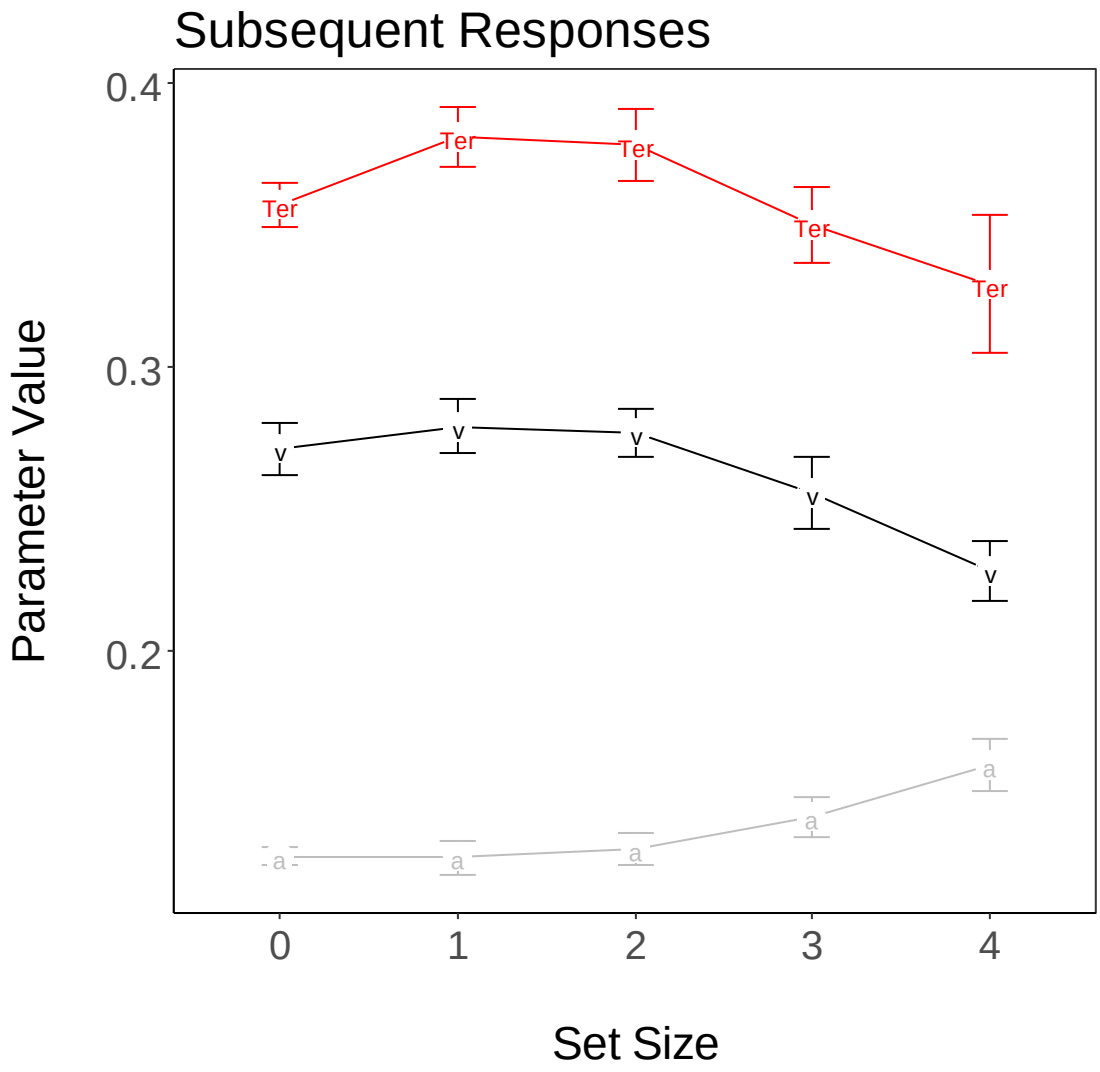

Figure 6. Drift rate (v), boundary separation (a), and time for encoding and executing the response (T_{er}) computed with the EZ diffusion model from the data of Experiment 2 are plotted against set size. Error bars represent standard errors for within-subjects designs.

A final analysis focused on the question whether articulatory rehearsal is gradually automatized, as suggested by Naveh-Benjamin and Jonides (1984). This gradual automatization would be reflected in a decrease of the set-size effect over processing positions. The second row of Figure 4 suggests such a decrease: The posterior of the set-size effect on RTs had a smaller mean for subsequent RTs compared to first RTs. However, this tendency most likely appeared because there was large uncertainty about the set-size effect of first RTs. Assessment of the interval covering 95% of the posterior density (hereafter 95%

highest density interval, HDI) is informative regarding parameter uncertainty (Kruschke, 2014). It can be seen in the lower panels of Figure 4 that the HDI was much larger for first RTs compared to subsequent RTs and last RTs. To formally test for a decline of the set-size effect over processing position, we ran a further Bayesian linear model on first, subsequent, and last RTs together, adding processing position (first, subsequent, or last) as further continuous predictor with a random slope. The BF for the main effect of processing position was 14.4, providing strong evidence for slower RTs in the first position compared to subsequent positions. Most importantly, the BF for the interaction of set size x processing position was 0.33, indicating that the Null hypothesis should be favored by a factor of 3. The slight evidence against the interaction between set size and processing position suggests that the attentional costs of articulatory rehearsal rather do not decline over time.

6.4.3. Discussion

Experiment 2 demonstrated in a Brown-Peterson task that articulatory rehearsal delays concurrent processing. This confirms that articulatory rehearsal demands central attention. Analyses with the EZ diffusion model showed that the attentional costs are largely due to an uptick in the attentional demand at set sizes 3 and especially 4. The increased attentional demand was mainly detected in the *v* parameter. This suggests that the effect of rehearsal at larger set sizes is not to postpone the response selection of the CRT task, which would have led to an increase of T_{er}. Rather, rehearsal and response selection appear to simultaneously share a limited central processing resource, which is assumed to slow down central processes such as response selection (Navon & Miller, 2002; Tombu & Jolicœur, 2003). An additional analysis showed that the attentional costs remained until the end of the processing period.

The result that up to two or three words could be rehearsed with negligible costs could be explained by assuming a phonological loop that is limited to the rehearsing of speech of about 2 seconds length (Baddeley, 2001). Once the limit is exceeded, an additional, attentionally demanding process (e.g., refreshing of the memory items) is recruited, as proposed by Vergauwe et al. (2014), which would explain the substantial increase in processing RTs from set size 3 on. However, the current data cast doubt on that explanation. A core assumption of the phonological loop model is that long words take longer to rehearse, so that fewer of them can be held in the loop (the word-length effect; Baddeley, 2012; Baddeley et al., 1975). This would have predicted an effect of the number of syllables on processing RTs Yet, the BF provided evidence against an effect of number of syllables. Therefore, we have to consider different explanations for the pronounced increase of processing RTs with set size 4. One tentative explanation is that participants start to make occasional rehearsal errors starting with set size 3 or 4. It has been shown that rehearsal errors attract attention (Phaf & Wolters, 1993), which would explain the attentional demand of rehearsal.

At first glance, the persistent attentional costs over the whole processing period, and the ambiguous evidence for a set-size effect when only considering set sizes 0 and 2, deviate from the results of Experiment 1. However, the comparison of processing RTs between Experiment 1 and Experiment 2 is not straightforward because the presentation of the CRT stimuli was probabilistic in Experiment 1. The first CRT stimulus could appear in any of the five segments. For a comparison between the two experiments we re-analyzed the data from Experiment 1, separating first and subsequent processing responses in each trial, regardless of which segment of the trial the first response occurred in. It can be seen in Figure 7 that the posteriors of the set-size effects for 0 vs. 2 words largely overlap between experiments. This shows that the results from the two experiments are in agreement with each other, and it

suggests that if there are any attentional costs of rehearsing two words, they must be fairly small, particularly during subsequent RTs.

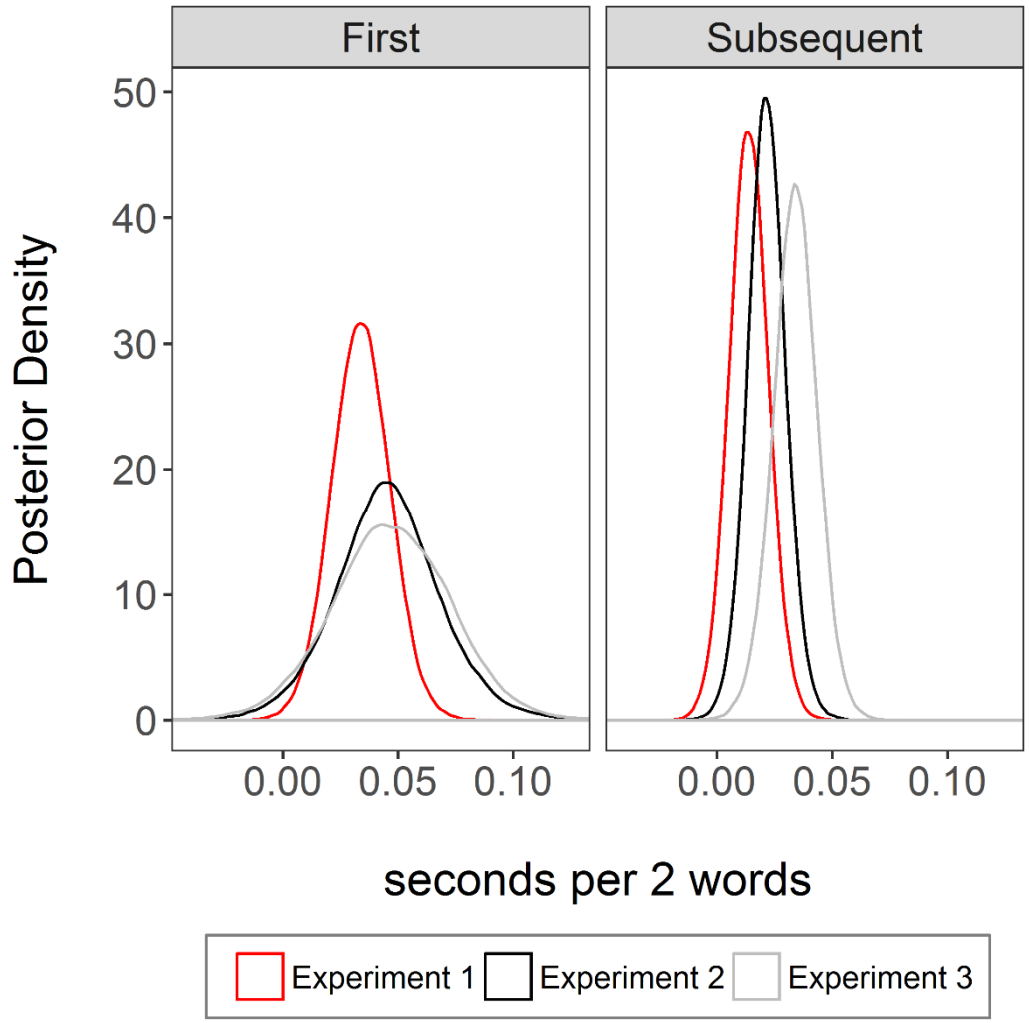

Figure 7. Posteriors for the difference between rehearsing 0 and 2 words from the data of Experiments 1, 2, and 3 (only from the articulatory rehearsal condition).

To conclude, Experiment 2 showed that articulatory rehearsal cannot be carried out without any central attentional costs. The costs persisted over a 10 s processing period and they started to magnify especially with set size 4. Analyses with the EZ diffusion model showed that the increase in processing RTs at larger set sizes is due to a reduction of drift rate, reflecting slower evidence accumulation.

6.5.Experiment 3

The goal of Experiment 3 was to assess again the attentional costs of rehearsal and elaboration, and in addition to assess whether participants engaged in refreshing, or in elaboration, when rehearsal was prevented through AS.

Vergauwe et al. (2014) investigated the set-size effect on processing RTs when people were asked to engage in AS. They assumed that AS forced participants to abandon articulatory rehearsal and to resort to refreshing, a more attentionally demanding maintenance process. They observed substantial set-size effects in the AS condition, which exceeded those in their silent condition. To firmly conclude that AS leads people to use a maintenance strategy more attentionally demanding than articulatory rehearsal, the attentional costs of overt articulatory rehearsal and of maintenance under AS have to be compared in the same experiment. Another possibility is that under AS, participants resort to elaboration rather than to refreshing. Self-report strategy data indicate that participants tend to use elaboration in about 1/3 of the trials in WM tasks (Bailey, Dunlosky, & Kane, 2008; Bailey et al., 2011; Dunlosky & Kane, 2007). Elaboration is also assumed to be attentionally demanding. Although Experiment 1 indicated that the attentional costs of elaboration were brief, this strategy may be more demanding when set size is increased, as was the case for articulatory rehearsal in Experiment 2. Experiment 3 was designed to assess the plausibility of these conjectures.

Experiment 3 combined features of Experiments 1 and 2. Participants were assigned to one of two groups (Elaboration or No Elaboration) and their memory for the word lists was assessed again in a surprise delayed-recognition test. There was an immediate memory test in all experimental conditions, and the processing task required responses throughout the processing period.

To investigate the attentional costs of maintenance processes other than articulatory rehearsal, participants were required to perform AS in half of the trials. Hence, participants in the No Elaboration group were instructed to either engage in overt articulatory rehearsal (AR condition), or to perform only AS (AS condition). Participants in the Elaboration group were instructed to either engage in elaboration silently (EL condition), or to perform AS in addition to elaboration (EL + AS condition). Finally, we also varied the concreteness and imageability of the words to be remembered to explore whether elaboration is more beneficial for concrete words with high imageability, which should be easier to elaborate, in comparison to abstract, difficult to imagine words.

This experimental design accomplished three goals. First, it allowed us to replicate the persistent costs of articulatory rehearsal, and its increase with set size. Second, it allowed us to assess whether the costs of elaboration would persist throughout the processing period with increased set sizes, and when a WM test was expected, unlike what we found in Experiment 1. Third, by assessing set-size effects under AS (with or without the instruction to use elaboration), we could measure the putative attentional costs of other maintenance processes participants may resort to in the absence of articulatory rehearsal. Under AS participants may do one of three things: (a) Refreshing, as postulated by Vergauwe et al. (2014). This should lead to a substantial set-size effect on first and subsequent RTs, as observed by Vergauwe and colleagues. (b) Elaboration – this should lead to attentional costs that are similar to the ones observed under the instruction to elaborate under AS. (c) Participants might simply do nothing apart from the requested AS. Overtly articulating irrelevant syllables such as "babibu" could entail its own attentional demand (yielding a main effect of AS), but this cost should not increase with memory load. Hence, in this case we expect no effect of memory set size on CRTs.

The design of Experiment 3 allowed us to test additional predictions arising from possibilities (a) – (c). A summary of the predictions is presented in Table 3. If people are elaborating spontaneously in the AS condition, performance in the AS condition should look similar to the EL + AS condition not only with regard to CRT dual-task costs, but also with regard to delayed recognition and WM performance. Concerning delayed recognition, we expect both elaboration (Craik & Tulving, 1975) and refreshing to increase delayed recognition (e.g., Grillon, Johnson, Krebs, & Huron, 2008; M. K. Johnson, Reeder, Raye, & Mitchell, 2002; M. R. Johnson et al., 2013; Loaiza, Duperreault, Rhodes, & McCabe, 2014; Loaiza & McCabe, 2012). Therefore, the elaboration conditions should lead to better delayed recognition performance than the AR condition. If people elaborate or engage in refreshing in the AS condition, that condition should also lead to improved delayed recognition, whereas if they do nothing, their delayed recognition should be no better than in the AR condition.

Table 3
Predictions for Performance in the AS Condition Depending on the Type of Maintenance Process Used by Participants.

	Hypothetical Maintenance Strategy		
Predicted Effect	(a) Refreshing	(b) Elaboration	(c) Nothing
Processing RTs:			
Set-size effect	First RTs and subsequent RTs	= EL + AS	Absent
Memory:			
Delayed test	> AR	> AR = EL + AS	= AR < EL + AS
WM test	?	= EL + AS	< EL + AS
Concreteness effect (immediate and delayed test)	= AR	= EL + AS	= AR

Note. AR = Articulatory Rehearsal; EL = Elaboration; AS = Articulatory Suppression.

Concerning WM performance, correlational self-report studies (Bailey et al., 2008; Bailey et al., 2011; Dunlosky & Kane, 2007) have suggested that elaboration improves WM performance. Our design allows a first experimental test of that conjecture: WM recall in the

EL condition should be better than in the AR condition. If participants in the AS condition engage in elaboration, it should improve WM performance too. At the same time, AS is known to decrease WM performance. Therefore, the beneficial effect of spontaneous elaboration in the AS condition can only be gauged by comparing WM performance in the AS condition to the EL+AS condition: If participants in the AS condition elaborate spontaneously, their WM performance should be comparable to that in the EL+AS condition, because the instruction to elaborate would make little difference to what people do spontaneously during AS. In contrast, if participants in the AS condition do nothing, their WM performance should be worse than in the EL+AS condition. No prediction can be made for the possibility that participants in the AS condition refresh, because we do not know whether refreshing leads to better or worse WM performance than elaboration.

Finally, we aimed to use the concreteness effect as a diagnostic tool to distinguish between different types of rehearsal. If concrete, highly imageable words benefit more from elaboration than abstract, poorly imageable words, participants in the elaboration conditions (EL and EL + AS) should show a larger concreteness effect in immediate and delayed memory tests compared to the AR condition. If this is the case, we can use the magnitude of the concreteness effect to diagnose whether people engage in elaboration in the AS condition.

6.5.1. Methods

Participants. Forty university students (30 women; *M* = 25 years old) were randomly assigned to one of two groups: No Elaboration (*n* = 20) or Elaboration (*n* = 20). One participant from the No Elaboration group mentioned after the experiment to have expected the delayed memory test. But given that this participant did not mention having used any additional strategy to remember the memoranda, the data were retained for analysis.

Materials. In the present experiment, we manipulated three variables within participants: set size (0, 2, or 4 words), word concreteness and imageability (concrete, highly imageable words vs. abstract, poorly imageable words), and articulatory suppression (without or with AS). For each condition created by the combination of these three variables, eight word lists were generated. Two sets of German words were compiled from the “Semantischer Atlas” data base (Schwibbe, n.d.). One set consisted of words with high ratings of concreteness and imageability (henceforth the concrete word pool, consisting of 96 items), whereas the other set consisted of words with low ratings on both dimensions (abstract word pool, also with 96 items). The word sets were equated for mean word length (mean = 7.8 characters) and frequency (mean log frequency among 4.5 million words = 4.9). For each participant, the memory lists were created by randomly sampling (without replacement) from the respective word pools.

The delayed recognition test was constructed in a similar fashion as described in Experiment 1 with the following exceptions. First, we always selected the first word of a word list as the cue and the second word of that wordlist as the correct alternative to control for serial position across lists with different lengths. There were 64 recognition trials in total – 32 trials for abstract words and 32 for concrete words – that were presented in a randomized order. The correct answer had to be selected amongst an intrusion word from another trial and two new words. Two pools of 64 new words were randomly sampled without replacement from the "Semantischer Atlas" with the only exception that they were not already used for the memory lists. New words for concrete recognition trials were taken from the first pool, new words for abstract recognition trials came from the second pool. The intrusion words selected from another trial could come from any serial position within that trial. Every intrusion probe appeared only once during the recognition test.

Procedure. Experiment 3 combined the procedures of Experiment 2 and Experiment 1. As in Experiment 1, participants were assigned to one of two groups that differed regarding the type of rehearsal instruction (No Elaboration or Elaboration). The sequence of events within a trial was exactly as described in Experiment 2 (see Figure 1b), with the following exceptions. First, set size was manipulated in a less fine-grained level. Across trials, participants were presented either with 0, 2, or 4 words. Moreover, unbeknownst to the participants, half of the lists consisted of concrete words, and the other half of abstract words. Third, the suppression condition (without or with EL instruction) was manipulated between blocks of trials.

Half of the participants in each group started the experiment with the AS condition (AS or EL + AS), and completed the rehearsal condition without AS (AR or EL) in the second half of the experiment. For the remaining participants, the order of these conditions was reversed. At the beginning of each condition, detailed instructions were displayed on the screen explaining the rehearsal or AS requirements. In the conditions requiring AS, participants were instructed to always articulate “babibu”, even when set size was zero. In contrast, in the conditions requiring articulatory rehearsal or elaboration (without AS), participants were instructed to remain silent in 0-words trials. The instruction was followed by three practice trials, one for each memory load (excluded from final analyses).

6.5.2. Results

Delayed Recognition. How effective were the different maintenance strategies in laying down an accessible LTM trace? Delayed recognition hit rates are plotted in Figure 8. They were log-odds transformed before the analysis. The upper part of Table 4 shows the BFs for the contrasts of interest. Concrete and highly imageable words were recalled much better than abstract, less imageable words, and the Elaboration group tended to show a larger

effect of concreteness/imageability than the No Elaboration group, confirming that participants in the Elaboration group followed the instruction to elaborate. Within the No Elaboration group, delayed recognition tended to be worse when participants rehearsed aloud than when they engaged in AS, replicating a result of Camos and Portrat (2015), but the effect of concreteness/imageability was comparable in these two conditions. The finding of better delayed memory in the AS condition suggests that participants used a different maintenance process than articulatory rehearsal in this condition; it is unclear whether this process is refreshing or elaboration as the concreteness effect was indistinguishable from both the AR and the EL + AS conditions. Finally, focusing on the two No-AS conditions, delayed recognition in the EL condition was better overall than in the AR condition, replicating Experiment 1, and showing that participants adhered to the elaboration instruction.

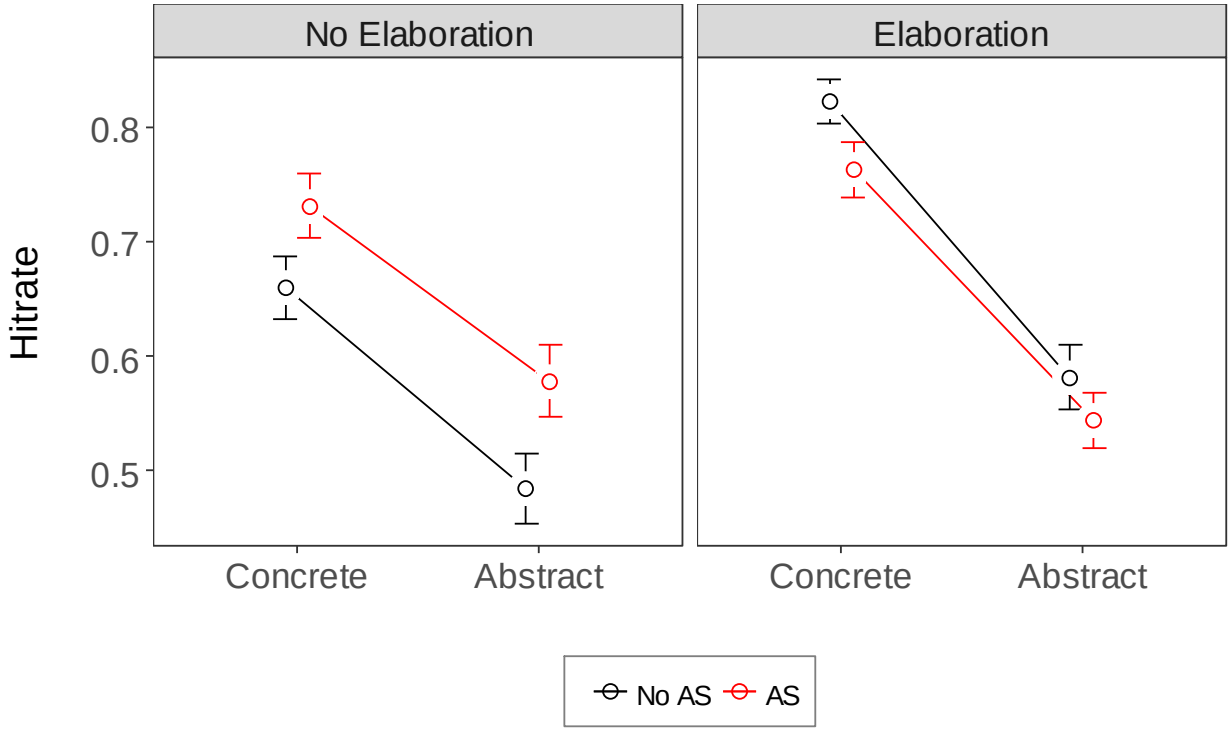

Figure 8. Average hit rates in the delayed recognition test for concrete and abstract words in the four rehearsal conditions in Experiment 3. Error bars represent standard errors for within-subjects designs.

Table 4. Upper panel: Bayes Factors for the contrasts of the model predicting proportion of delayed recognition hit rates and a t-test comparing the EL and the AR conditions. Lower panel: Bayes Factors for the contrasts of (a) the model predicting serial recall accuracy in the WM recall test (proportion correct) for all eight data points in Figure 9 and (b) of the model predicting only serial recall accuracy in the conditions including an AS instruction.

Table 4. Upper panel: Bayes Factors for the contrasts of the model predicting proportion of delayed recognition hit rates and a t-test comparing the EL and the AR conditions. Lower panel: Bayes Factors for the contrasts of (a) the model predicting serial recall accuracy in the WM recall test (proportion correct) for all eight data points in Figure 9 and (b) of the model predicting only serial recall accuracy in the conditions including an AS instruction.

Memory Test	Effect	BF
Delayed Recognition Test		
All Data	Word Concreteness	9.80E+11
	Group (Elaboration or not)	1
	Word Concreteness x Group	2.2
No-Elaboration Group	Articulatory Rehearsal vs. Articulatory Suppression	2.4
	Word Concreteness x Articulatory Rehearsal vs. Articulatory Suppression	0.5
No AS Conditions	T-Test: Elaboration vs. Articulatory Rehearsal	9.66
AS Conditions	Word Concreteness x Articulatory Suppression vs. Elaboration + Articulatory Suppression	0.89
Working Memory Test		
All Data	Word Concreteness	8.00E+10
	Articulatory Suppression	7.40E+19
	Group (Elaboration or not)	1.4
	Word Concreteness x Articulatory Suppression	0.7
	Word Concreteness x Group	2.7
	Articulatory Suppression x Group	0.4
	3-way Interaction	0.7
AS Conditions	Word Concreteness x Articulatory Suppression vs. Elaboration + Articulatory Suppression	2.8

WM Recall. Our analysis of WM recall was limited to set-size 4, because recall at set-size 2 was very close to ceiling. For brevity, we concentrate on those contrasts that could be informative about which maintenance process participants might have used. Accuracy data were log-odds transformed before the analysis (see Figure 9 for the data and the lower part of Table 4 for the BFs).

Concrete words were remembered better than abstract words, and AS decreased serial recall accuracy. The elaboration instruction tended to improve serial recall accuracy overall, and it tended to increase the effect of concreteness/imageability. When only comparing the two conditions including AS, instructing people to elaborate still tended to increase the concreteness effect. This result suggests that participants in the EL + AS condition elaborated the words more than participants in the AS condition.

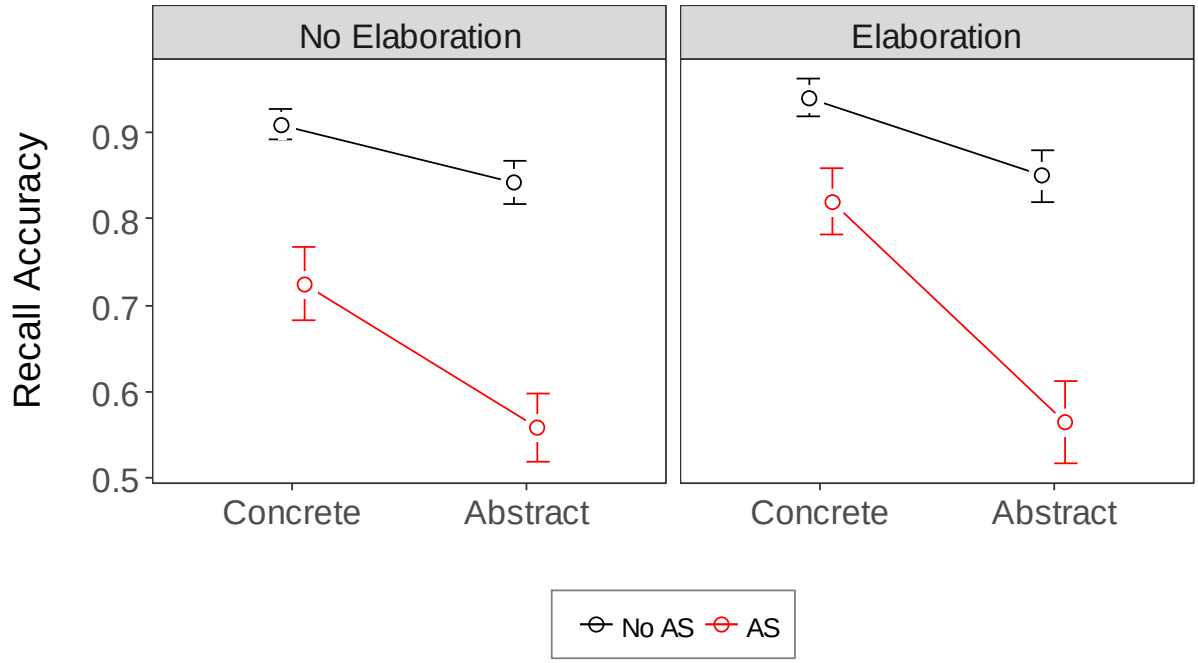

Figure 9. Mean serial recall accuracy in the WM test as a function of word concreteness in the four rehearsal conditions in Experiment 3. Error bars represent standard errors for within-subjects designs.

Processing Task. The main dependent variable of interest in this task was RT. Processing accuracy was overall high ($M \sim 95\%$) and there was no evidence for an effect of any of the manipulated variables in this measure (analysis not reported here).

For the analysis of RTs, we only included trials in which participants succeeded in recalling all memoranda in their correct serial position in the WM task. This led to the exclusion of 15.47% of all trials (representing 25.98% of all trials with set size > 0). RTs in the retained trials were further trimmed by removing incorrect responses in the processing task (4.51% of the remaining responses), and by dropping RTs that exceeded 5 s (0.04%). Processing RTs are plotted against processing position for every combination of rehearsal instruction and suppression condition in Figure 10.

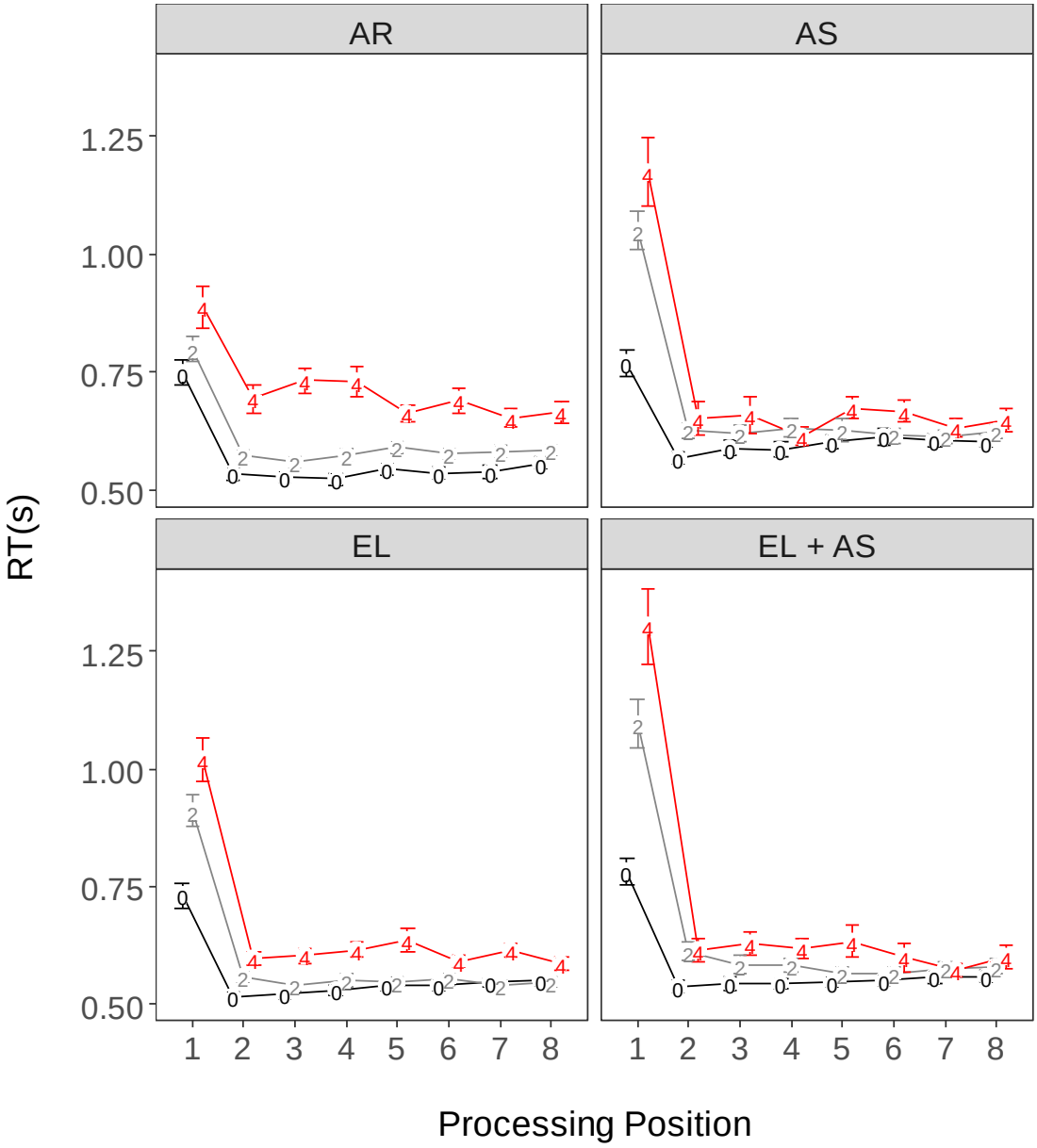

Figure 10. RTs as a function of processing position. Panels represent the four different rehearsal conditions in Experiment 3. The numbers in the graphs represent the three set-size conditions that are also plotted in different colors. Error bars represent standard errors for within-subjects designs.

Our main interest was in estimating the attentional costs of each rehearsal strategy as reflected in the slope of the set-size effect. We excluded word concreteness from these analyses because an initial analysis showed that word concreteness had no influence on processing RTs (BFs = 3.45, 11.1, and 6.3 for the Null for first, subsequent, and last RTs,

respectively). In order to estimate the posterior of the set-size slope in each condition and for each measure of interest (first, subsequent, and last RTs), we entered set size (as a numerical predictor) and rehearsal condition as predictors (including random slopes for these variables) in a hierarchical Bayesian regression model (henceforth called the full model) that was run via JAGS. Specifically, we estimated the set-size slope in each condition and computed the BF for this effect against the Null. In addition, we computed the pairwise comparisons of the set-size slopes between conditions. For all analyses we report the BFs and DICs (see Table 5). The graphical representation of the full model and the respective priors are shown in Figure 11.

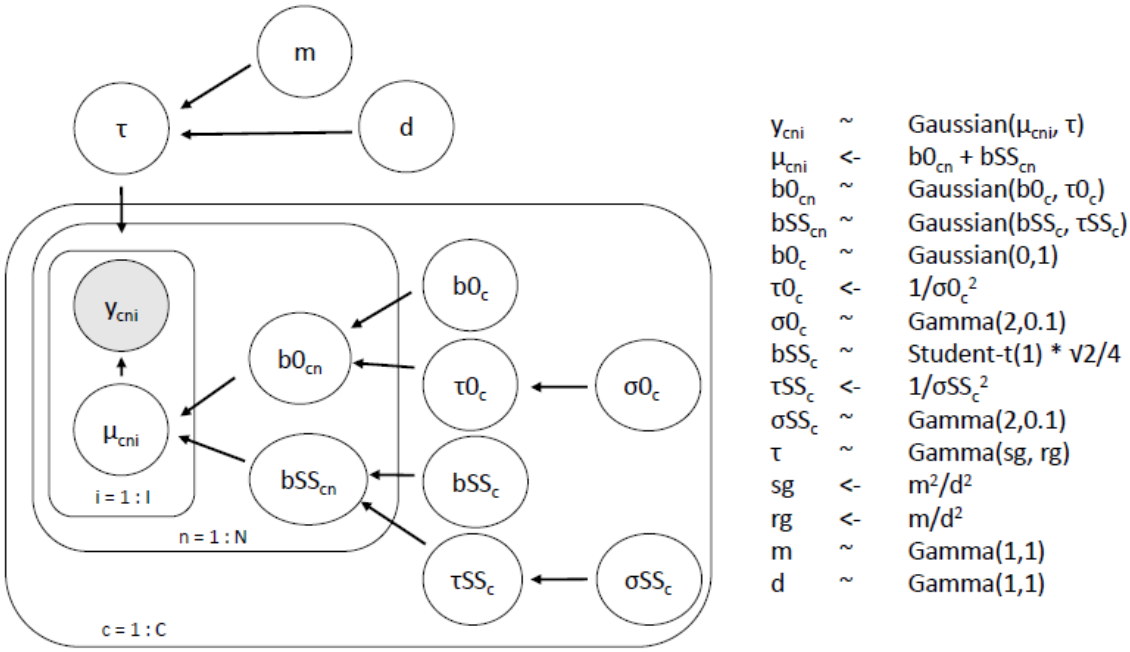

Figure 11. Graphical representation of the regression models run in JAGS used to predict RTs for each condition in Experiment 3. The circle shaded in gray represents the data, the circles without shading represent variables to be estimated. The c-plate indicates independent repetitions over C conditions, the n-plate over N participants, and the i-plate over I trials. SS = set size.

Table 5
BFs and DICs of the set-size effect in the four conditions (upper table) and of the pairwise comparisons between the set-size slopes in the four conditions.

	Processing Position					
	First		Subsequent		Last	
Comparison	BF	ΔDIC	BF	ΔDIC	BF	ΔDIC
Set-size effect in each condition						
AR	7.6	-63.43	$1.7*10^8$	-273.98	183'933	-72.10
AS	$2.3*10^9$	-127.81	22	-76.75	0.21	2.15
EL	6049.00	-143.84	107'313	-113.83	4.6	-16.38
EL +AS	9898.00	-229.59	3.6	-68.59	0.18	-3.43
Pairwise comparison of the set-size effect between conditions						
AR vs. EL	0.35	0.80	8.4	-2.91	1.4	-1.53
AR vs. AS	52	-30.26	16.00	-69.93	11	-10.05
AR vs. EL +AS	11	-4.50	11	-0.99	19	-5.02
EL vs. AS	0.47	0.68	0.086	1.61	0.1	1.86
EL vs. EL +AS	0.72	-49.02	0.1	4.60	0.11	5.43
AS vs. EL +AS	0.16	-2.11	0.091	1.14	0.066	3.36

Note. ΔDICs in the upper half of the table represent DIC (set size model) – DIC (Null model). Hence, values below zero favor the set-size model. ΔDICs in the lower half of the table represent the DIC difference between a model that allows the set-size slopes to differ between the two conditions to be compared and a model in which the difference is 0. Hence, values below zero favor the model in which the slopes differ between conditions. Gray shading indicates divergence between the BF and the DIC measures.

Empirical RT means are plotted as a function of set size in Figure 12, and the means of the set-size posteriors are shown in Table 6. We observed credible CRT costs in the AR condition that persisted until the end of the processing period, replicating the finding of Experiment 2. The set-size posteriors for first and subsequent RTs only using set sizes 0 and 2 are shown in Figure 7 for comparison with Experiments 1 and 2. AS and EL+AS delayed processing even more than AR in the beginning of processing. However, the costs in these conditions decreased over the processing period and were no longer distinguishable from zero in the last processing episode. The CRT costs in the EL condition were somewhat in between: They were neither clearly distinguishable from those in the AR condition nor from those in the AS and EL+AS conditions. The posterior means of the set-size slopes (Table 6) show that the CRT costs in the EL condition were persistent throughout the retention interval. However,

compared to the AR condition they were numerically larger in the beginning of processing but numerically smaller in the end of processing.

Table 6
Posterior mean slopes of the set-size effect (ms per word) for each of the four rehearsal conditions in Experiment 3.

	Processing Position		
	First	**Subsequent**	**Last**
AR	37	35	32
AS	109	13	7
EL	70	16	13
EL +AS	127	13	6

This pattern does not align nicely with the idea that only elaboration has persistent central attentional costs but articulatory rehearsal does not (Naveh-Benjamin & Jonides, 1984). We evaluated the hypothesis of Naveh-Benjamin and Jonides in an additional analysis, focusing only on the AR and EL conditions, and testing whether the CRT cost (i.e., the set-size effect) decreases from first to last RTs more for elaboration than for articulatory rehearsal. RTs were predicted by set size, condition (AR vs. EL), and processing position (first or last RTs) and all their interactions. Random effects were specified for all main effects as well as the intercept. The three-way interaction is shown in Figure 13 as the two-way interaction between condition and position contrast on the set-size slopes. The BF for the three-way interaction was 20.28, supporting the hypothesis that the costs in the EL condition decreased more over time than in the AR condition. This trend is the opposite of the one reported by Naveh-Benjamin and Jonides (1984).

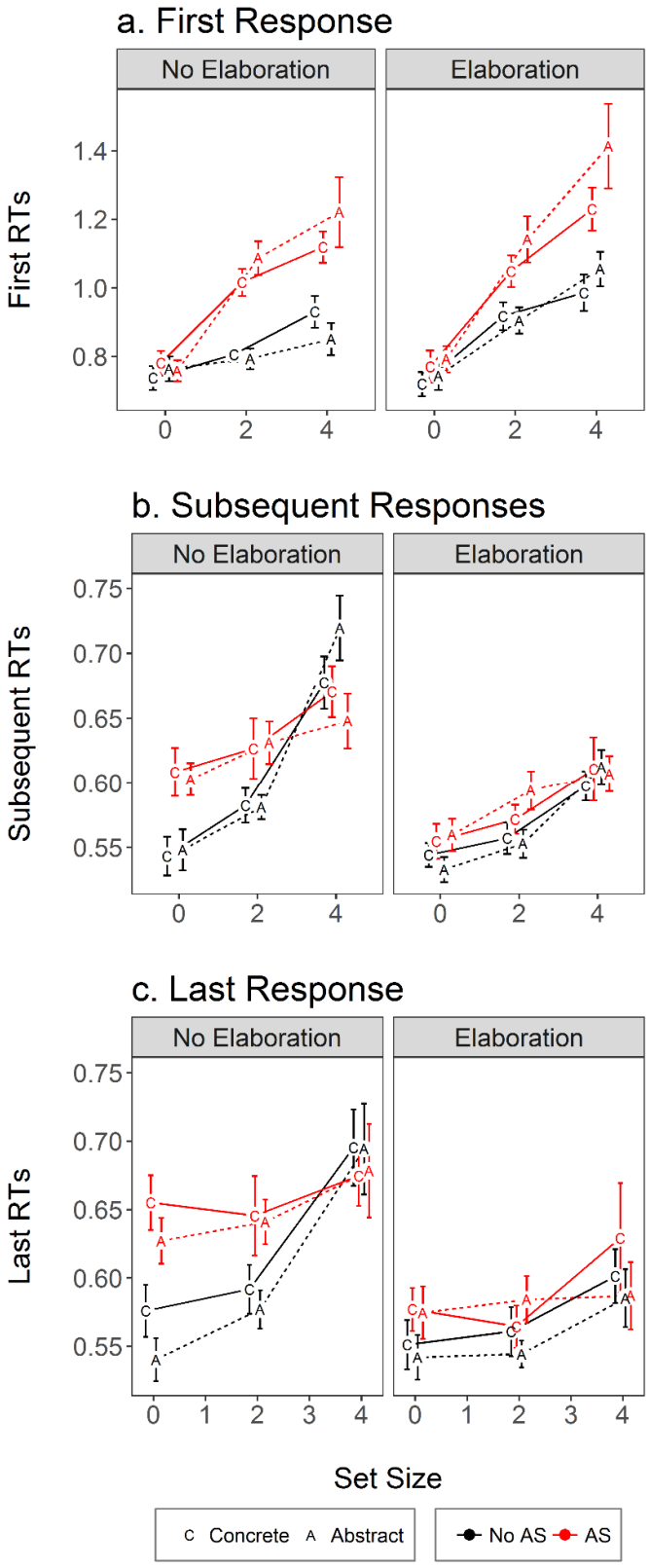

Figure 12. Empirical RT means as a function of set size for first (panel a), subsequent (panel b), and last (panel c) RTs in Experiment 3. Note the different scales on the y-axes in the panels. Error bars represent standard errors for within-subjects designs.

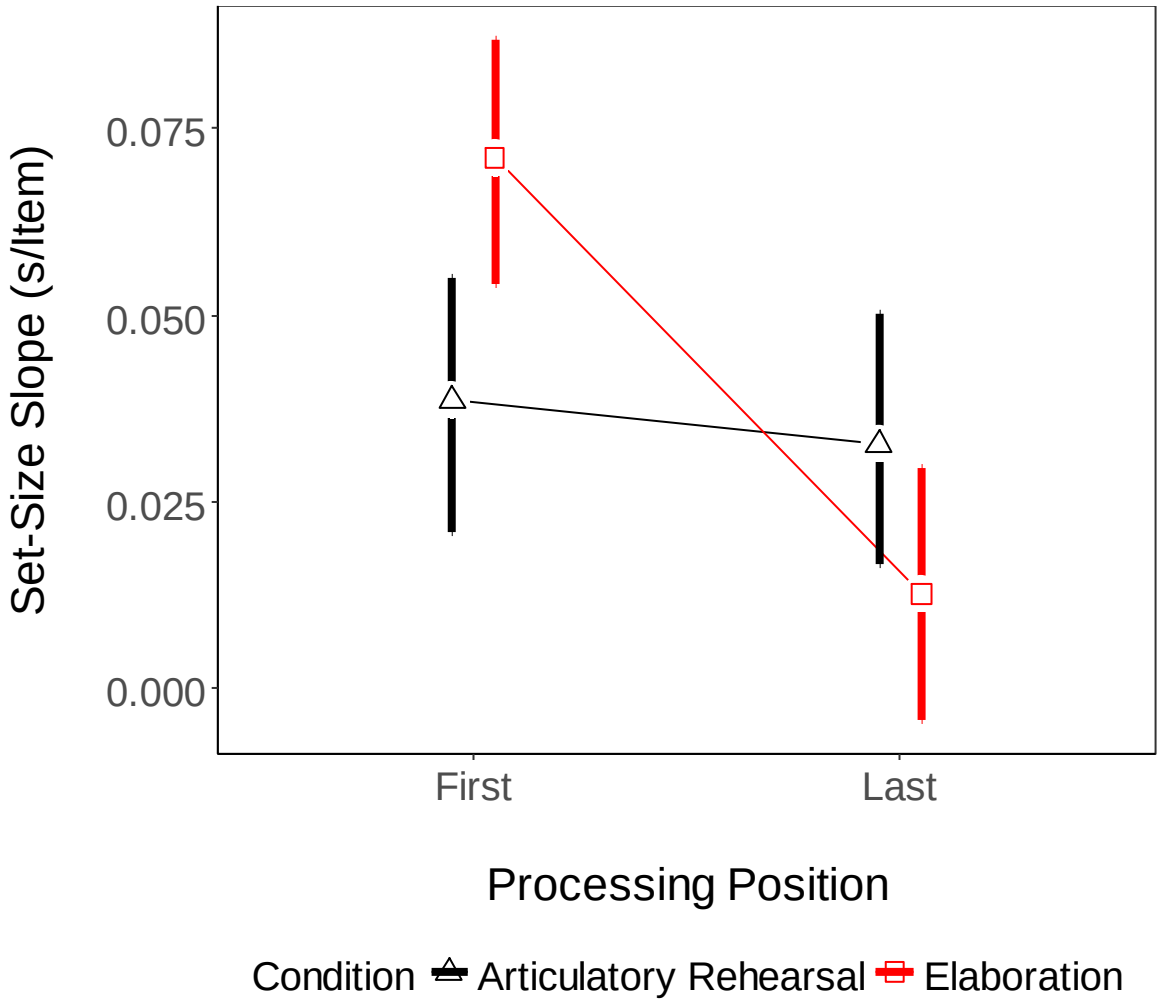

Figure 13. Posterior means and 95% HDIs of the three-way interaction of rehearsal instruction x position contrast x set size in Experiment 3. The three-way interaction is presented as a two-way interaction of rehearsal instruction x position contrast on set-size slopes.

As in Experiment 2, the increase in processing RTs in the AR condition appeared to be driven by set size 4. We analyzed subsequent RTs with the EZ diffusion model to test whether the increase in processing RTs is actually due to central attention. The parameters are separately plotted for each condition against set size in Figure 14. An initial analysis using the data from all four conditions showed that T_{er} was unaffected by condition and set size (all BFs < 1). Therefore, the following analyses focused on *v*, which is diagnostic for a central-attention demand of rehearsal that slows down response selection in the CRT

(whereas parameter *a* is diagnostic for an effect of a speed-accuracy trade-off, which is not of primary interest here). We assessed the set-size effects separately in the four conditions with Bayesian linear models using the BayesFactor package. Set size was entered as categorical predictor with random slopes. We additionally compared the set-size effect between the AR and AS conditions due to the following reason. Based on the analyses of processing RTs we have to consider that in the AS condition people do not use any attentionally demanding process to remember the words. When AS only prevents articulatory rehearsal, we can compare the AR and AS conditions to estimate the attentional demand of articulatory rehearsal while controlling for load effects. The BFs for the analyses are shown in Table 7. The results showed that there was strong evidence for a set-size effect on *v* in the AR condition, driven by the lower drift rate at set size 4. This replicates the finding from Experiment 2 that the increase of processing RTs due to articulatory rehearsal is only attributable to central attention at set size 4. There was substantial evidence for a set-size effect in the EL condition, also due to a lower drift rate at set size 4. The evidence was against a set-size effect in the AS and EL + AS conditions. The set size x AR vs. AS interaction received some support, reflecting the tendency for a larger drop in *v* between set sizes 2 and 4 in the AR condition than in the AS condition.

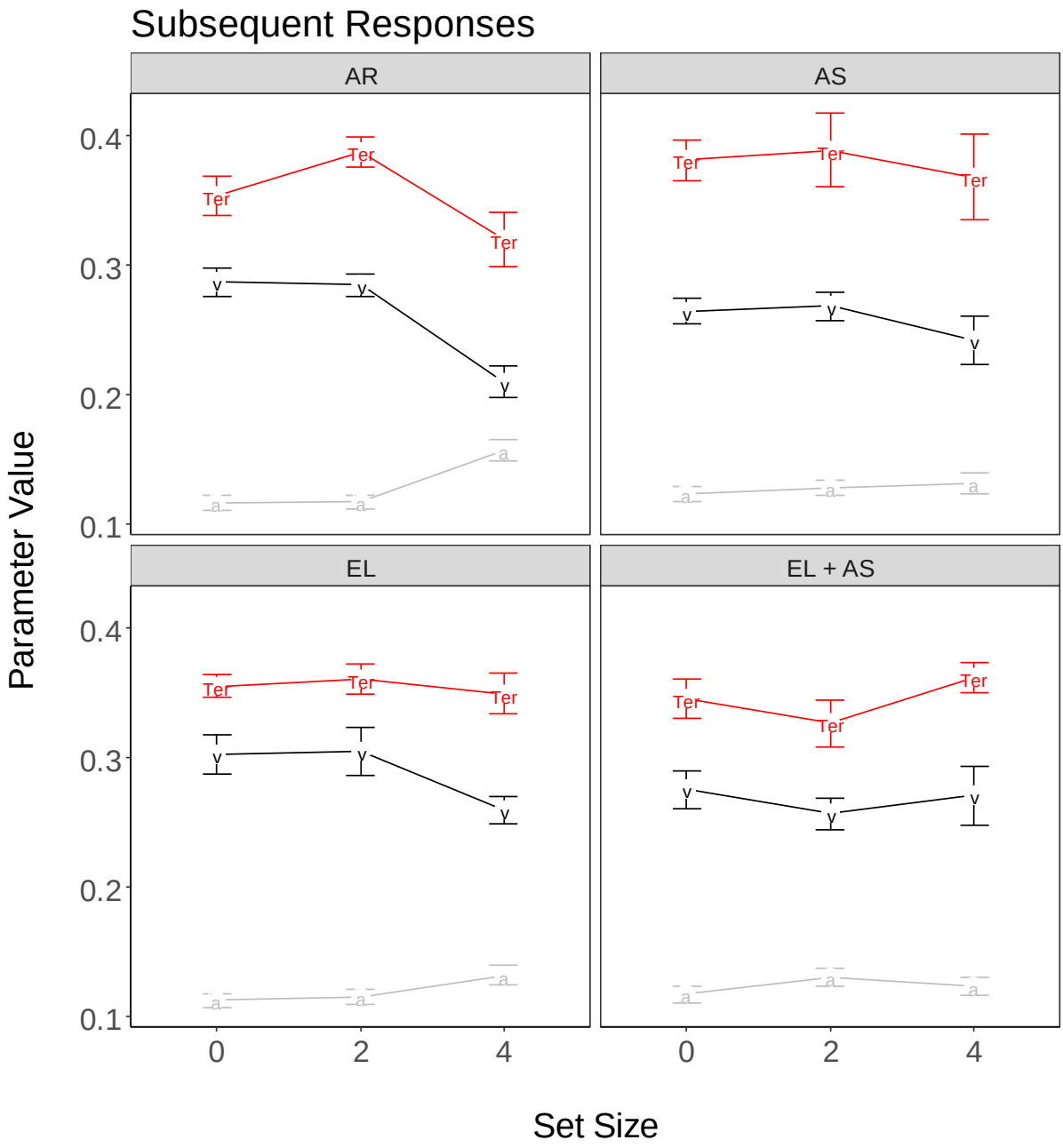

Figure 14. Drift rate (v), boundary separation (a), and time for encoding and executing the response (T_{er}) computed with the EZ diffusion model applied to the subsequent processing RTs in Experiment 3. Estimates are plotted as a function of set-size separately for each rehearsal conditions (different panels). Error bars represent standard errors for within-subjects designs.

Table 7. BFs for the effects of interest using drift rate *v* as dependent variable in Experiment 3.

Condition	Effect	BF
Articulatory Rehearsal	Set Size	14’402
Articulatory Suppression	Set Size	0.69
Elaboration	Set Size	5.4
EL + Articulatory Suppression	Set Size	0.32
Articulatory Rehearsal vs. Articulatory Suppression	Set Size x Condition	4.8

6.5.3. Discussion

Experiment 3 shed light on the attentional costs of all three forms of rehearsal: Articulatory rehearsal, elaboration, and refreshing. First, we discuss the findings regarding elaboration. Then, we consider what the results from the AS and the AR condition tell us about the most elusive of the three processes, refreshing, and about articulatory rehearsal.

The attentional demand of elaboration peaked at the first processing RT and then dropped quickly. Our results suggest that elaboration requires a substantial amount of central attention initially, probably for generating an interactive image of the words. Once this is accomplished, the central attentional demand of elaboration drops to a lower level. This could mean that, once an interactive image has been created, maintaining it involves a smaller demand on central attention than creating the image.

What can we learn from Experiment 3 about refreshing? Vergauwe et al. (2014) assumed that AS motivates participants to engage in refreshing of a verbal memory list. This should lead to a substantial effect of memory set size on RTs. In our AS condition we observed a large set-size effect only on the first RTs of each processing period, whereas the set-size effect on subsequent RTs dropped to a low level. This was also observed in the EL + AS condition. Compared to the AR condition, the AS condition was attentionally more demanding only in the initial phase of maintenance. The opposite was true for subsequent and

last RTs, in which the demands on central attention tended to be larger in the AR condition than in the AS condition. This contradicts the assumption that participants continuously engage in refreshing during AS throughout the retention interval.

We considered the possibility that participants in the AS condition may use elaboration, rather than refreshing. If that was the case, the AS condition should yield comparable effects on CRTs and on memory performance as observed for the EL + AS condition. The evidence for this prediction is mixed. On the positive side we observed similar set-size slopes on CRTs, and indistinguishable concreteness effects in delayed recognition. On the negative side the concreteness effect in immediate recall tended to be larger with EL + AS than with AS alone. The AS condition tended to result in better long-term memory than the AR condition, a finding replicating a similar observation by Camos and Portrat (2015), and implying that participants in the AS condition did something to improve memory. It appears plausible that participants in the AS condition engage in some attention-demanding process – which may be refreshing or elaboration – briefly at the beginning of the processing interval, which is reflected in their large set-size effect on first RTs, and which caused a modest improvement of long-term memory. In sum, our results pertaining to the AS condition are best explained by the simple assumption that, after a brief initial effort, participants in this condition engage in no attention-demanding maintenance process at all.

Experiment 3 showed again that instructing people to rehearse aloud delays processing. As in Experiment 2, the increase of processing RTs with set size (a) persisted until the end of the processing period and (b) was only attributable to central attention (i.e., slowed response selection) with set size 4. Vergauwe et al. (2014) hypothesized that people use refreshing in addition to rehearsal when more than three words have to be rehearsed. This would explain the central attentional cost at set size 4 because refreshing is attentionally demanding. A similar explanation would be that the participants started to use elaboration

when set size was larger than three, because elaboration is also attentionally demanding. The results of Experiment 3 render neither of these hypotheses plausible. If the sustained costs of AR were due to the use of refreshing, it is unclear why people would use it throughout the retention interval only in the AR condition (in which they could still rely on articulatory rehearsal to maintain information) but not in the AS condition (in which the option to rely on articulatory rehearsal was blocked). Furthermore, the continuous use of refreshing should have led to superior delayed recognition performance in the AR condition than in the AS condition; instead we observed the reversed pattern. Similarly, we can exclude that people used elaboration in addition to articulatory rehearsal because delayed recall in the AR condition was far worse than in the EL condition.

We conclude that the attentional cost of articulatory rehearsal at set size 4 is not due to simultaneous refreshing or elaboration. Rather, the cost arises from articulatory rehearsal itself. Moreover, the analysis with the EZ diffusion model shows that the effect is on drift rate, implying a cost on response selection, a processing stage requiring central attention. One reason why rehearsing four words (but not two words) engages central attention could be that rehearsing a larger number of words induces rehearsal errors, which attract central attention for detecting the error and correcting the speech plan (Phaf & Wolters, 1993).

6.6.General Discussion

We examined the central attentional costs of articulatory rehearsal and elaboration, and assessed the plausibility that people resort to refreshing when articulatory rehearsal is blocked (i.e., under AS). Prior investigations did not allow conclusions about the central attentional costs of these rehearsal mechanisms for several reasons. The studies of Guttentag (1984) and Naveh-Benjamin and Jonides (1984) used a processing task that did not require response selection. Response selection is critical for measuring central attentional costs (Pashler, 1994). Vergauwe et al. (2014) used a processing task requiring response selection, but participants were not explicitly instructed to perform articulatory rehearsal or refreshing. Therefore, whether and how participants actually rehearsed or refreshed could not be ascertained. We overcame these limitations by using a CRT to measure central attentional costs, and by instructing participants to engage in overt cumulative articulatory rehearsal, or in elaboration. We also applied a formal RT model that allows to disentangle central attentional effects from speed-accuracy trade-off effects.

The primary finding is that articulatory rehearsal clearly requires central attention throughout the retention interval when at least four words have to be rehearsed. The fact that rehearsal requires attention confirms the conclusion of Naveh-Benjamin and Jonides (1984). In contrast to these authors, we show that the attentional cost persists for at least 10 s when the memory set consists of more than two or three words.

The fact that about three words can be rehearsed with very little costs could be explained by assuming a phonological loop, which has a capacity of about 2 s of speech (Baddeley, 2001). Building on this assumption, Vergauwe et al. (2014) argued that the attentional demands increase from set size 4 because people start to use refreshing once the capacity of the phonological loop is exceeded. Similarly, it could be that people start to use elaboration in addition to articulatory rehearsal at higher set sizes. Experiment 3 showed that

none of these two possibilities is plausible. Furthermore, Experiment 2 showed that it is unlikely that the attentional costs of articulatory rehearsal are due to the overload of the phonological loop. A phonological loop account would have predicted larger attentional costs for disyllabic than for monosyllabic words because longer words take more time to be rehearsed, exceeding the phonological-loop capacity already at smaller set sizes (Baddeley et al., 1975). No such effect was observed in Experiment 2, and Vergauwe et al. (2014; Experiment 5) also did not find an effect of the number of syllables on CRTs. A tentative explanation for the cost of articulatory rehearsal is that people start to make more rehearsal errors from set size 4 on. Rehearsal errors are known to attract attention (Phaf & Wolters, 1993), and this could explain the increased attentional demand.

The second finding was that elaboration requires central attention, but this demand decreases quickly after all memoranda have been presented. This finding is not in agreement with the hypothesis that elaboration constantly occupies the central attentional bottleneck to the same extent (Naveh-Benjamin & Jonides, 1984). We suggest that participants instructed to elaborate create a mental image of the memoranda, a process which is highly attentionally demanding. Once the mental image is created, maintaining it is less attentionally demanding. Briefly engaging in elaboration seemed to be sufficient for a modest improvement of immediate and delayed memory compared to articulatory rehearsal, in particular for concrete, highly imageable words, which arguably are easier to elaborate.

Our conclusions regarding the third rehearsal strategy, refreshing, are only based on the results of the AS condition in Experiment 3. Participants in the AS condition appeared to engage in an attentionally demanding strategy briefly after encoding: The set-size slopes on first processing RTs were larger with AS than with articulatory rehearsal. The shallow set-size effect on subsequent RTs, and the failure to find any set-size effect on drift rate, imply that memory maintenance in the AS condition demands no central attention. Moreover, the

set-size effect completely vanished for the last processing RTs. These results imply that asking participants to engage in AS does not induce a persistent attentionally demanding strategy.

In many regards – in particular the time course of the attentional demand – the AS condition looked similar to a condition with instructed elaboration (EL + AS). Both conditions resulted in better long-term memory than articulatory rehearsal. One noticeable difference between AS with and without elaboration was that instructed elaboration tended to specifically improve WM for concrete words. That effect was small and received only modest statistical support. Therefore, at present we cannot decide whether the spontaneous maintenance strategies participants engage in during AS and the EL+AS are qualitatively different (i.e., refreshing as opposed to elaboration) or merely quantitatively different (namely, less systematic elaboration across trials in the AS condition compared to the EL+AS condition).

To conclude, we showed that articulatory rehearsal delays concurrent processing. The RT costs were most noticeable with size 4 and were persistent to the end of a 10 s processing period. In contrast, elaboration imposed a large cost on central attention primarily briefly after the memoranda were presented, and these costs were reduced thereafter. Finally, preventing articulatory rehearsal through AS does not induce participants to engage continuously in an attention-demanding strategy such as refreshing or elaboration. These findings require a re-conceptualization of the costs of different rehearsal processes in WM models in three regards. First, because articulatory rehearsal does to some extent require central attention, it cannot be assumed to operate in parallel with refreshing without costs, contrary to Camos et al. (2009). Rehearsing and refreshing in parallel should be possible when people only have to remember two or three words. However, so far no one has claimed that several rehearsal mechanisms are required to remember only two or three words;

experiments investigating WM usually involve larger set sizes. Second, contrary to Naveh-Benjamin and Jonides (1984), elaboration does not persistently demand central attention to a larger degree than other maintenance processes. Third, contrary to Vergauwe et al. (2014), people do not spontaneously engage in persistent refreshing, or any other central-attention demanding strategy, when maintaining a list of verbal items while engaging in articulatory suppression.

7. Estimating Bayes Factors for Linear Models with Random Slopes on Continuous Predictors

Mirko Thalmann[ab], Marcel Niklaus[a], and Klaus Oberauer[a]

[a] University of Zurich

[b] University of New South Wales

Submission status:

Submitted for publication

Authors‘ contributions:

Mirko Thalmann: Programming and testing the R package, programming and analyzing the simulation studies, writing the manuscript

Marcel Niklaus: Discussing the results, testing of the R package and the simulation studies, commenting on the manuscript

Klaus Oberauer: Supervising and discussing the project and its implementation, commenting on the manuscript

Acknowledgement

This research was partly supported by a Swiss National Science Foundation Doc.Mobility grant to MT (#168257).

7.1. Abstract

Using mixed-effects models and Bayesian statistics has been advocated by statisticians in recent years. Mixed-effects models allow researchers to adequately account for the structure in the data. Bayesian statistics – in contrast to frequentist statistics – can state the evidence in favor of or against an effect of interest. For frequentist statistical methods, it is known that mixed models can lead to serious over-estimation of evidence in favor of an effect (i.e., inflated Type-I error rate) when models fail to include individual differences in the effect sizes of predictors ("random slopes") that are actually present in the data. Here, we show through simulation that the same problem exists for Bayesian mixed models. Yet, at present there is no easy-to-use application that allows for the estimation of Bayes Factors for mixed models with random slopes on continuous predictors. Here, we close this gap by introducing a new R package called BayesRS. We tested its functionality in four simulation studies. They show that BayesRS offers a reliable and valid tool to compute Bayes Factors. BayesRS also allows users to account for correlations between random effects. In a fifth simulation study we show, however, that doing so leads to slight underestimation of the evidence in favor of an actually present effect. We only recommend modeling correlations between random effects when they are of primary interest and when sample size is large enough. BayesRS is available under https://cran.r-project.org/web/packages/BayesRS/, R code for all simulations is available under https://osf.io/nse5x/?view_only=b9a7caccd26a4764a084de3b8d459388

7.2. Introduction

Researcher X is planning a study to test whether a new drug has a beneficial effect on reading scores. To that end, X applies the drug on several days with different dosages to 40 adult participants. She also wants to be able to assess the strength of evidence for the null hypothesis because it is possible that the drug does not have an effect at all. Although X's research question seems straightforward, answering it adequately has been made possible only recently with the development of tools to analyze data in the Bayesian framework (Carpenter et al., 2016; Lunn, Thomas, Best, & Spiegelhalter, 2000; Plummer, 2003) and with the development of tools to analyze nested data (e.g., Pinheiro, Bates, DebRoy, & Sarkar, 2014). In order to adequately test her hypothesis, X has to consider two points in her statistical analysis.

First, because X wants to generalize her results to the general population of adults she has to use a statistical model that accounts for systematic variabilities between the units of observation in the sample. People will likely differ in their average reading ability. Hence, X has to add a random intercept (i.e., individual differences in the intercept) to the model. People may also differ in their sensitivity to the dosage manipulation of the drug. Hence, X has to add a random slope (i.e., individual differences in the size of a predictor's effect) on the dosage predictor. Second, she wants to use Bayesian statistics because then she can obtain evidence for the Null and test whether the Null is more likely than the Alternative.

Unfortunately, there is no easy-to-use statistical tool available at the moment that can tell X how strong the evidence in favor of or against a relationship between dosage of the drug and reading scores is. Currently available statistics packages do not allow us to compute the evidence for continuous predictors (such as dosage) with associated random slopes within a Bayesian framework. This is problematic because it has been shown that not accounting for true random slopes leads to massive over-reporting of effects when they are actually not

present (i.e., Type 1 errors; Barr et al., 2013); in the framework of null hypothesis significance testing (NHST). Here, we show that the same problem exists also within the Bayesian framework. To help researchers address this problem, we introduce a new package in the R statistical environment (R Core Team, 2017) that is able to compute Bayes Factors (BFs) for continuous variables with associated random slopes.

In the following, we will introduce the basic principles of linear mixed-effects models and of the Bayesian statistical framework. This will clarify why the combination of these two represents a great asset for psychological research. After that, we will introduce the new R package, BayesRS, and test its functionality in five simulation studies. To foreshadow, BayesRS provides a viable new tool for researchers to analyze their data.

Bayesian Statistics. At the core of the Bayesian framework is the idea to express the believability of a proposition as a probability distribution (Kruschke, 2014). That is, probability is not defined as the expected frequency of an event, resulting from an imaginary infinite number of samples from the reference set (Neyman, 1977), but rather reflects a subjective belief that a proposition is true. Therefore, we can assign probabilities not only to events but also to hypotheses and theories. After having observed some data speaking to a proposition, we update our subjective belief in it based on these data according to Bayes' rule. Statistical inference in the Bayesian framework proceeds through three steps: First, specification of a prior probability distribution (hereafter prior) of a hypothesis; second, calculating the likelihood of the hypothesis given the data; and third, computing a posterior probability distribution (hereafter posterior) by updating the prior with the likelihood. Using the posterior for inference about the value of a parameter in a statistical model uses all of the available information (i.e., relevant prior knowledge and the data). For unimodal, symmetric posteriors it is convenient to summarize them with their mean and an interval, for example

the interval covering the most credible 95 % of its density (95 % highest density interval, HDI, e.g. Kruschke, 2014).

A major advantage of Bayesian statistics over classical NHST is that it allows for the comparison of the relative plausibility of competing hypotheses (H) given some data (D). This comparison is tightly linked to Bayes' Theorem (Bayes, Price, & Canton, 1763)

$$P(H|D) \propto P(D|H) * P(H)$$

It states that the posterior P(H|D) is proportional to the likelihood P(D|H) multiplied by the prior P(H). Bayes Theorem can also be expressed in terms of odds, that is, ratios of probabilities pertaining to the two competing hypotheses:

$$\frac{P(D|H_1)}{P(D|H_2)}.$$

This equation shows that we should update the prior odds by the likelihood ratio to obtain the posterior odds. The likelihood ratio is also known as the Bayes Factor (e.g., Jeffreys, 1935; Kass & Raftery, 1995). The inverse of the BF reflects the BF in favor of the competing hypothesis. Kass and Raftery (1995) suggest some loose guidelines how BFs might be interpreted. A BF in the range of 1-3.2 is considered to be "not worth more than a bare mention", a BF ranging from 3.2-10 as "substantial", a BF from 10-100 as "strong", and a BF > 100 as "decisive" evidence in favor of a hypothesis.

The Bayesian framework offers solutions to problems that cannot be addressed easily within the framework of NHST (Wagenmakers, 2007; Hubbard, 2004) and thus represents a powerful alternative to conventional inference based on *p* values.

Linear Mixed-Effects Models. In a seminal paper, Clark (1973) pointed out a problem he termed "The Language-as-Fixed-Effect Fallacy". He argued that experimental researchers often have a small pool of items in their experiments but nevertheless would like to

generalize their results to a class of items in general. By neglecting this fact in their statistical analyses they are likely to commit Type 1 errors. That is, they report an effect to be true in the broader class of items, even though there is actually no effect in the population of items but only in the specific sample of items. Hence, accounting for the fact that items are a random effect, and not a fixed effect, is of great importance.

To say that an effect is fixed means that the true value of the effect is the same across all observation units of a set (e.g., a set of subjects, items, studies) and differences between values across observations units are only due to sampling noise (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2010). In contrast, to say that an effect is random means that the true values of the effect differ across observation units. Specifically, it is assumed that the true values across units reflect a random draw from a larger population that is described by a distribution of values, often a normal distribution. When the dependent variable is modeled as a linear combination of independent variables and measurement noise, and includes fixed as well as random effects, the resulting class of models is called linear mixed-effects models. Mixed-effects models have only become popular in psychological research in the last few years, because statistical software that allows their computation has become available only recently (e.g., Pinheiro et al., 2014).

Although mixed-effects models represent a powerful tool, the increased complexity in model structure is accompanied by increased degrees of freedom in the model building process. For instance, the question how one should set up the random-effects structure in a given set of data is still a matter of debate. One view is that analysts should use a fit criterion and apply the random-effects structure that best fits the specific set of data (Baayen, Davidson, & Bates, 2008). Another view is that the maximal random-effects structure should be used that is justified by the design of the experiment . That is, even though there may be no substantial differences in an experimental effect across the different observation units,

researchers should nevertheless include a random slope for this effect in their model. Barr and colleagues (2013) showed via simulation that a model including random slopes generalizes better to an underlying population in terms of Type 1 and Type 2 errors.

Random effects for continuous predictors. Although the debate regarding the appropriate random-effects structure continues (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017), the point put forward by Clark (1973) – that models that do not include a random effect that is actually present are prone to increased Type 1 errors – is still valid. Including random slopes in a model can therefore be an important step when analyzing experimental data. The BayesFactor package (Morey & Rouder, 2014), which computes BFs for mixed models, allows users to specify random intercepts and random slopes for categorical predictors (i.e., variables on a nominal scale), as are often used in experimental research. However, at present, there is no statistical software that allows users to readily compute BFs for models that include random slopes on continuous predictors, that is, predictors with multiple levels on an ordinal or even an interval scale of measurement. This is problematic as experimental researchers often include variables in their experiments that are continuous – as in the example of researcher X who was interested in modeling the effect of the dosage of a drug on reading ability. Here, we present a new R package, BayesRS, that closes this gap. We test the reliability and validity of the BFs computed with the package (Simulations 1 to 3), and investigate the consequences of including or not including random slopes, and their correlations, in mixed-effects models (Simulations 4 and 5).

7.3. Method

The BayesRS package allows the specification of linear mixed-effects models in the Bayesian framework with maximally two levels in the hierarchy. Here, we describe the model structure for the full hierarchical model, in which the coefficients of all predictors are

estimated for all observation units of a set. The dependent variable and the continuous predictors are entered as z-standardized values into the model. Categorical predictors with *n* levels are entered as *n*-1 simple coded contrasts. Therefore, fixed effects or group distributions of the mean effects reflect standardized regression coefficients. An important feature is the use of default priors that we put on mean effects size, as will be seen in the following. Every single z-standardized data point yz_i is modeled according to

$$yz_i \sim \mathcal{N}(M_i, T) \quad (1)$$

where $\mathcal{N}$ is short for the probability density function of the normal distribution. The linear regression structure is on

$$M_i = \sum_{p=0}^{P} \sum_{j=1}^{J_p} x_{ij}^{(p)} \beta_j^{(p)}, \quad (2)$$

and precision (the inverse of the variance) is defined as

$$T \sim \Gamma(a,b). \quad (3)$$

We follow Gelman, Carlin, Stern, and Rubin (2014) regarding the notation in batches. The first sum in Equation (2) runs over batches; each batch stands for one predictor. Every p^{th} batch represents a set of J_p random regression coefficients belonging to a specific predictor (e.g., a categorical predictor, a continuous predictor, or an interaction). Because the model includes random slopes, each observation unit has its own beta coefficient for each predictor; hence each batch represents *J* beta coefficients. Across all batches, the $\beta_j^{(p)}$ coefficients are entries of a P x J matrix, where P and J reflect the number of predictors and the number of observation units in all sets, respectively. The intercepts are represented in row number “0” of the matrix. That is, each observation unit has its own intercept reflecting the mean of that specific observation unit across all levels of the predictors. We assume the individual by-

subjects beta coefficients to be normally distributed according to a group distribution with mean μ_p and condition precision $\frac{1}{\sigma^2}_p$

$$\beta_j^{(p)} \sim \mathcal{N}\ (\mu_p, \frac{1}{\sigma^2}_p) \quad (4)$$

and further

$$\mu_p \sim T(0, df = 1) \times \text{Scale}_j, \quad (5)$$

$$\sigma_p \sim \Gamma(1,0.04). \quad (6)$$

Here, T stands for the Student's *t*-distribution that is scaled according to a scaling factor Scale_j. We follow previous work by placing a Cauchy distribution (i.e., a Student's *t*-distribution with 1 degree of freedom) as a default prior on μ_p (Rouder, Morey, Speckman, & Province, 2012). The logic of placing a zero-centered Cauchy distribution on the standardized effects is to place somewhat more prior probability to larger effects than, for example, a normal distribution would. We use scaling factors of $\sqrt{2/4}$ and ½ for continuous and categorical predictors, respectively. They reflect default priors that are called "medium" in other software (Rouder et al., 2012), in contrast to a "wide" prior that puts even more prior probability on larger effects.

The package also allows users to estimate only one single β coefficient for a predictor *p*, in which case the p^{th} row of the P x J matrix drops out and only μ_p is estimated for predictor *p*. In the following, we will refer to these mean parameters as fixed effects to facilitate comparison with the classical mixed-effects parlance.

Overview of the package. The BayesRS package is built in the R statistical environment (R Core Team, 2017). We use JAGS (Plummer, 2003) to sample from the posteriors of model parameters. JAGS itself is accessed from R via the rjags package

(Plummer, 2013). For processing and plotting of the data we make use of several other R packages that are mentioned in Appendix A.

The package takes as input dependent variables that are normally distributed. The dependent variable and all the continuous independent variables are automatically z-standardized. This means that the modeling is done in standardized space, and therefore all regression coefficients of continuous variables reflect standardized β coefficients. Hence, a β posterior has to be retransformed into the original scales when the effect size in the original scale (i.e., the b parameter) is of interest. Categorical predictors are entered into the model via simple coding.

The user accesses the package with its interface function called modelrun. The modelrun function does all the processing before and after the model is run in JAGS. It transforms the input data into the required format and calls the translation function modeltext that takes a description of the data structure as an input and writes it as a text file in JAGS language. After that, the modelrun function hands over the data and the written text file to JAGS, which then samples from the posteriors. After that, the modelrun function computes the 95% HDIs and the BFs of the parameters of interest. These are finally plotted in a figure, and the BF is returned as a vector. The modelrun function requires three obligatory arguments and has nine additional arguments that can be changed, but are otherwise set to default values.

The first three obligatory arguments reflect the data – which have to be handed over as a data frame in the long format –, the dependent variable with its name written as a string, and the data structure. The latter is a separate data frame that lists each predictor in a row, and defines for each whether it is continuous or categorical, and for what observation unit a random slope is required or not. That is, the package allows also for several, crossed-random effects. An additional optional argument allows users to put random slopes on interactions of

predictors. Three further optional arguments allow for the variation of the number of MCMC steps for adaptation and burn-in, and the number of MCMC steps that are actually saved. Further, it is possible to obtain convergence statistics of the main parameters in the model. Convergence of the model is a prerequisite for the BFs to be interpretable.

Another argument allows the user to specify which random effects within an observation unit should be modeled as correlated. The package can model correlations between pairs of variables, and correlations between more than two variables. When both is done at once, the variables within the two correlation structures cannot overlap (e.g., one can specify a pairwise correlation between predictors A and B, and additionally, the full correlation matrix between predictors C, D, and E).

One optional argument allows users to save the chains of the random intercepts and random slopes. Two final options are to determine (a) whether a figure of the 95% HDIs and the BFs of the main parameters should be plotted and (b) whether the deviance information criterion of a model (DIC, Spiegelhalter, Best, Carlin, & van der Linde, 2002) should be returned. Two examples of how the modelrun function can be used are available in the help file of the R package (https://cran.r-project.org/web/packages/BayesRS/).

Computation of Bayes Factors. A main feature of BayesRS is its ability to compute BFs for models including continuous predictors. We compute BFs with the Savage-Dickey density ratio (e.g., Dickey & Lientz, 1970). The Savage-Dickey density ratio is an approximation of the BF for nested model comparison. Following Wagenmakers, Lodewyckx, Kuriyal, and Grasman, (2010) we assume the null hypothesis to be a point null hypothesis, stating that the effect of a predictor in question is exactly zero. The alternative hypothesis states that the effect is non-zero. We can obtain the BF in favor of the alternative hypothesis by comparing a model that estimates the effect in question as a free parameter to an otherwise identical model in which it is fixed at zero. The BF is the ratio of the density of

the free parameter's prior to the density of its posterior in the alternative model, evaluated at the parameter value to which it is set in the null model (i.e., at zero).

Let us assume that researcher X has collected all data from her study and tests whether application of the drug increases reading test scores. Her null hypothesis is that the drug has no effect on reading scores. The alternative hypothesis allows effect size to vary freely in the model. In the alternative model, she has the default Cauchy prior on standardized effect size. After running the model, she observes a posterior with a mean of 0.5 and a standard deviation of 0.2. She obtains the BF for the alternative hypothesis by dividing the height of the prior by the height of the posterior at zero. An illustration is shown in Figure 1.

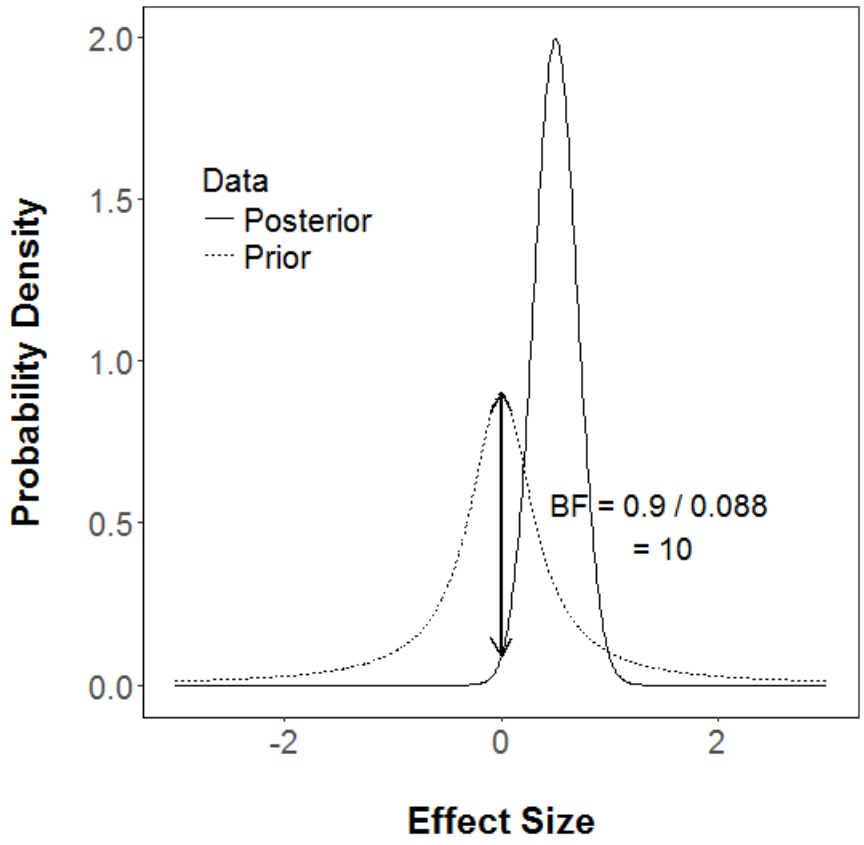

Figure 1. Demonstration of how the BF is calculated with the Savage-Dickey density ratio test. The null hypothesis is that effect size = 0. Therefore, the BF is the ratio of the density of the prior and the density of the posterior at zero.

Computationally, we use the approach by Wetzels, Raaijmakers, Jakab, and Wagenmakers, (2009) to estimate the density of the posterior at zero. That is, we compute the mean and the standard deviation from the MCMC chain of an effect of interest. Then, we estimate the density of the posterior at zero by computing the density of a normal distribution at zero with the computed mean and standard deviation. This approach takes advantage of the Bayesian central limit theorem that states that under general regularity conditions posterior distributions tend to be normally distributed when the number of observations becomes large (Wetzels et al., 2009).

7.4. Results and Discussion

In the following, we present five simulation studies that tested the functionality of the BayesRS package. We asked (1) whether the instantiation of the package leads to a reliable estimation of a BF given one set of data, (2) whether the computed BF on a fixed-effect model is comparable to an analytic solution provided by Gaussian quadrature, (3) whether the BF for the fixed effect of a continuous predictor approaches the BF of a Bayesian *t*-test on the true individual *b* values in the sample, (4) whether not accounting for random slopes when they exist in the data leads to increased potential Type 1 errors, and finally (5) whether not accounting for correlations between random slopes when they exist in the data leads to increased potential Type 1 errors.

Simulation Study 1. The first study aimed to test the reliability of the instantiation of the Savage-Dickey density ratio in the BayesRS package. Data were simulated from an experiment with one continuous independent variable with five values, and one random variable (e.g., subject). We independently varied the mean effect size of the continuous variable (0, 0.2, 0.5, 0.8), the number of subjects in the experiment (20, 40), and the number

of saved MCMC steps (50’000, 100’000). We sampled ten observations per design cell of the simulated experiment for each subject. The individual by-subject intercepts were generated according to a normal distribution with mean 0 and standard deviation 1. The by-subject slopes of the continuous variable were generated according to a normal distribution with the before-mentioned mean effect sizes, and standard deviation 1. Because we only generated one set of data for every design cell of the design of the simulation study, we forced the simulated by-subject slopes to exactly reflect the desired properties in this first study. On every generated data set we computed the BF 50 times with the modelrun function.

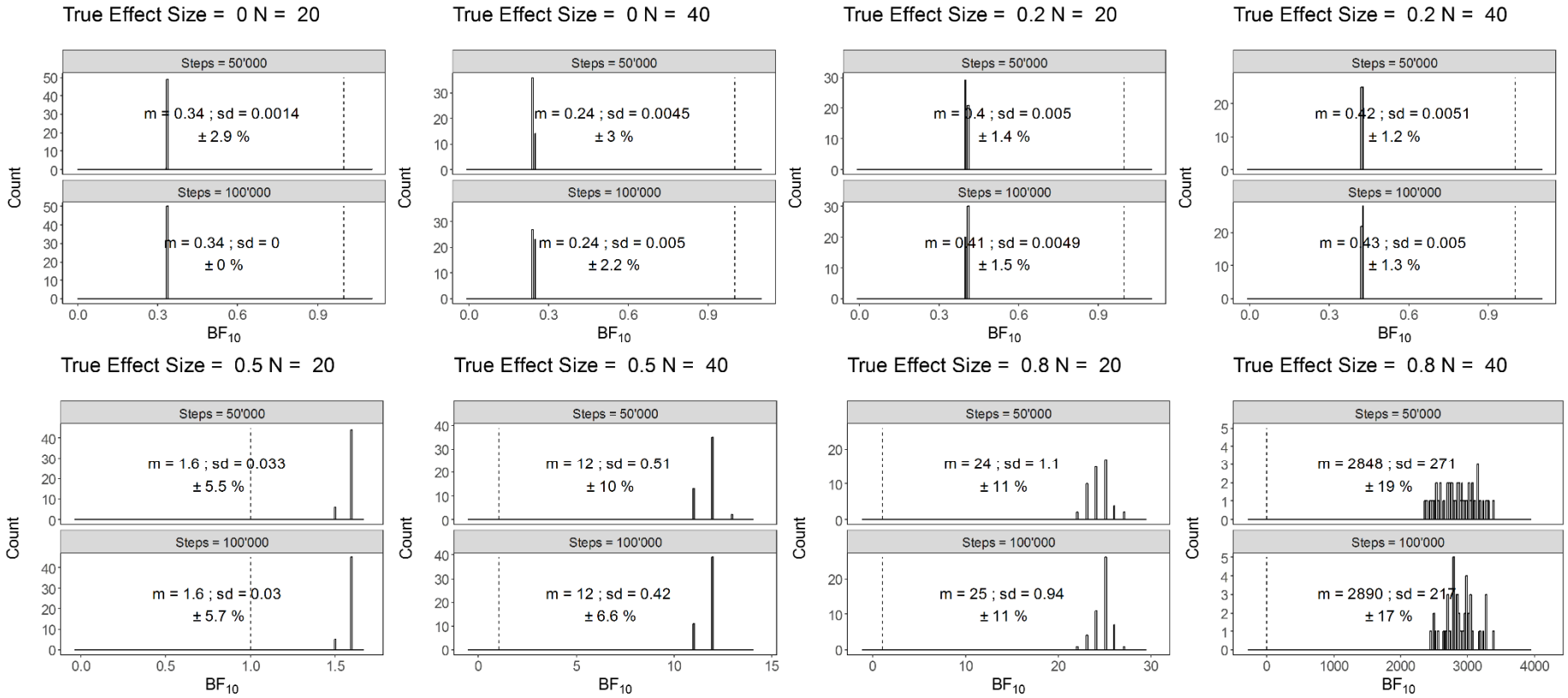

Figure 2. Distributions of BFs within the design cells of Study 1. We also show mean (m), standard deviation (sd), and the deviation of the most distant BF compared to the mean BF (deviation/mean BF*100) out of the 50 computed BFs per cell. The dashed line shows a BF of 1 representing a situation in which H_0 and H_1 are equally likely.

It can be seen in the results displayed in Figure 2 that BFs below 1 are estimated with high reliability. For example, a BF with a mean of 0.34, as in the first panel from the left in the upper row, can be estimated with perfect reliability with 100’000 MCMC samples, and a standard deviation of 0.0014 with 50’000 MCMC samples. When the BF grows larger, as with a mean of 12 in the second panel from the left of the lower row in Figure 2, it is sometimes estimated as 11 with 100’000 samples, and occasionally as 11 or (twice) as 13 with 50’000 samples. Although this represents a standard deviation of about 0.5, it is only a small, negligible loss in precision and would not change the interpretation of the results, i.e., that there is strong evidence in favor of the alternative hypothesis. When BFs grow even larger, they become more variable. In the most extreme case in the lower row in the fourth panel, the maximal and the minimal BFs were estimated to be 3392 and 2360, respectively, when 50’000 MCMC steps were saved. The variation with 100’000 samples was somewhat smaller. Although this shows substantial variation in the BF, it would not change the interpretation of the results. That is, in all cases, there is decisive evidence in favor of the alternative hypothesis.

To summarize, the estimation of densities that are further away from the mean of the posterior distribution is more variable. This is a consequence of the Savage-Dickey estimation of the BF, which relies on precise measurement of the posterior density at zero. When the mean of the posterior distribution grows larger, its density estimation at 0, being further out in the tail of the distribution, becomes less precise. Because the density is already very low, even absolutely small changes in its estimation have a relatively large influence on the estimation of the BF. Using more MCMC samples tended to attenuate the imprecision slightly. However, in none of the presented cases this variation would have led to changes in the interpretation of the data. Hence, we conclude that the ability to discriminate between the Null and the Alternative of the new R package is good enough.

Simulation Study 2. The aim of the second study was to test the validity of BayesRS, that is, whether the instantiation of how the BF is estimated yields comparable results to an existing benchmark. Specifically, we generated data with a continuous predictor without a random effect, and compared the BF when computed with the new package to the BF computed with the BayesFactor package (Morey & Rouder, 2015). The latter analytically computes the BF via Gaussian quadrature, and therefore we will call it the "true" BF in the following. We generated data from an experiment as in Study 1, but without random slopes on the continuous variable. Effect size was varied from .05 to .1 to .3, and data were randomly sampled without any constraints. The results are pooled across the different effect sizes and plotted in Figure 3 on a natural logarithm (ln) scale. Values to the right and to the left of zero on the x axis show BFs in favor of the Alternative and in favor of the Null, respectively. Values above the red line show cases in which the BF for the Alternative is larger when computed with BayesRS compared to the true BF, and values below the red line show cases in which the BF for the Null is larger when computed with BayesRS compared to the true BF.

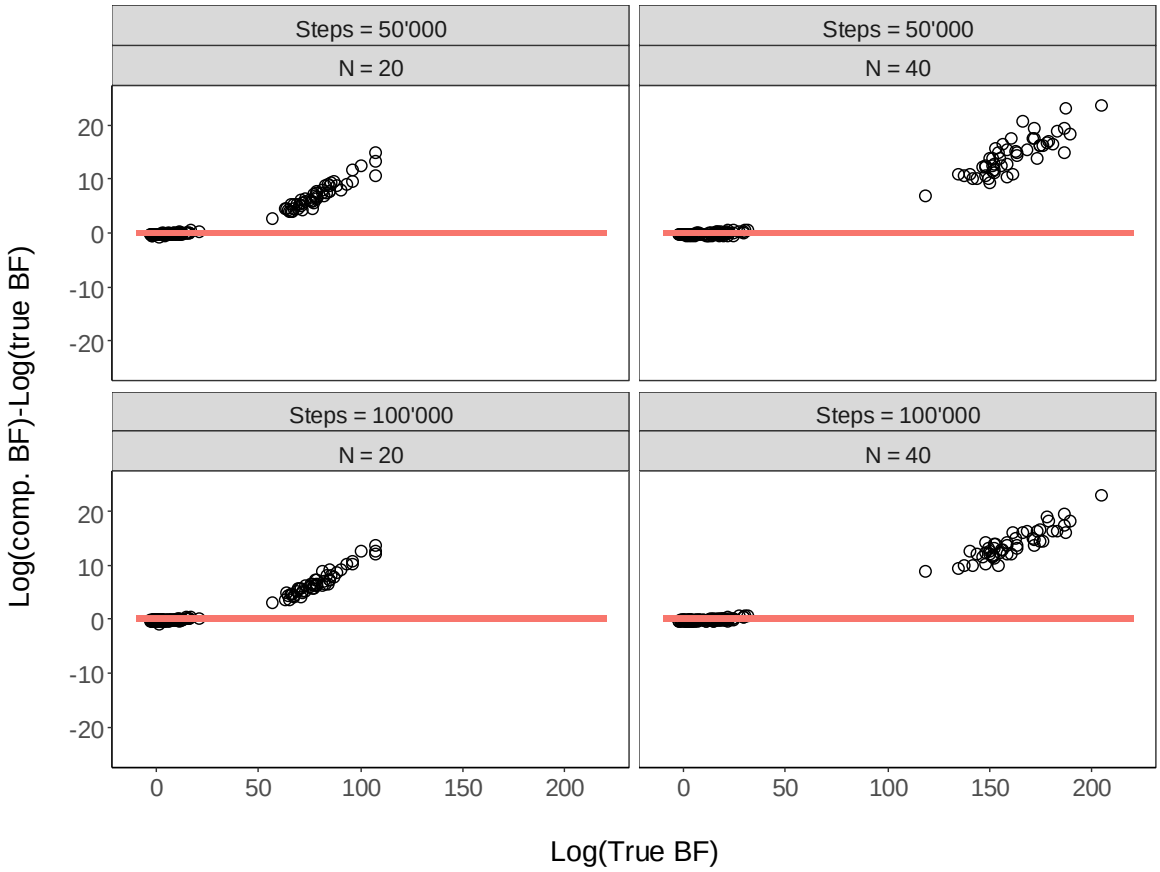

Figure 3. Deviation of the BF estimated with BayesRS from the true BF computed with the BayesFactor package, plotted against the true BF. Data are pooled across effect size and units are in Log space. A value on the red line indicates that the BF computed with BayesRS is the same as the true BF.

Figure 3 shows that the BF calculated with the Savage-Dickey density ratio represents the true BF acceptably. The approximation of the true BF is slightly better and slightly less variable with 100’000 MCMC samples than with 50’000 MCMC samples. Nevertheless, even with 100’000 MCMC samples, the estimation of the BF does not exactly match the true BF. The distortions are however in a desirable way. For true BFs in favor of the Alternative the computed BFs systematically overestimate the evidence in favor of the Alternative. The reverse is true as well: For true BFs in favor of the Null, the computed BFs systematically overestimate the

evidence in favor of the Null. When we zoom in to log(true BFs) between -2.3 and 2.3 – values that do not express strong evidence in favor of either hypothesis –, Figure 4 shows that the Savage-Dickey density ratio slightly overestimates the evidence in favor of the Null in all of these cases. In the most extreme case (see the outlier in Figure 4), our implementation underestimates a true BF of about 4.7 in favor of the Alternative as only about 1.9 in favor of the Alternative. Arguably, this underestimation is not severe. Overall, BayesRS yields comparable results to an analytical solution and therefore passes the second test.

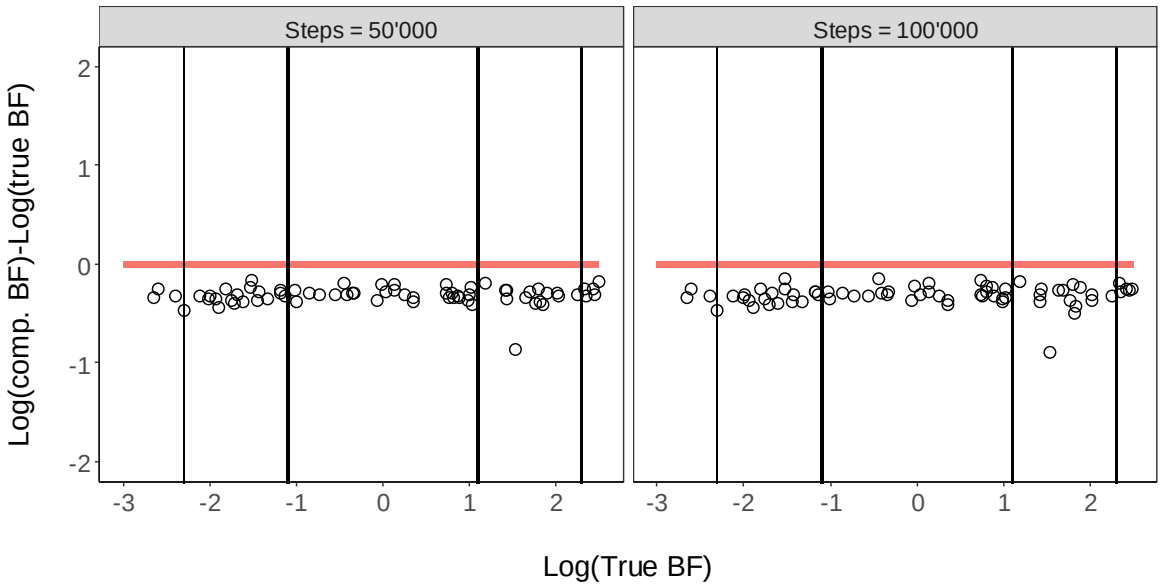

Figure 4. Deviation of the BF computed with the BayesRS package from the true BF, plotted against the true BF in Log space, zooming in on BFs that are within a range of relatively high uncertainty, where small errors in estimating the BF would matter most according to the classification of Kass and Raftery (1995). Note: The bold vertical lines correspond to BFs of 3.2 in favor of the Null (negative values) or the Alternative (positive values). The thin vertical lines correspond to BFs of 10 in favor of the Null (negative values) or the Alternative (positive values).

Simulation Study 3. The third study tested whether the BF of a fixed effect of a continuous predictor with random slopes asymptotically approaches the BF of a Bayesian *t*-test from the BayesFactor package on the known, true *b* values of individual observation units that were used

to generate the data set. In the following, we refer to the latter again as the “true” BF. The data were again simulated from a similar experiment as in Study 1, except that we varied the number of subjects between 10, 25, 50, and 100, and varied the number of observations per design cell between 3, 6, and 9. With these values we aimed to get more BFs that are closer to 1, which are crucial according to the classification by Kass and Raftery (1995) to decide whether an effect is not worth mentioning ($0.3125 < BF < 3.2$) or considered to be substantial ($BF > 3.2$ or $BF < 0.3125$) or strong ($BFs > 10$ or $BF < 1/10$). In addition, we fixed the number of MCMC steps to 100’000 because of slightly higher reliability compared to 50’000 steps. The results are depicted in Figure 5 and show a similar pattern as in Study 2. That is, BFs between -2.3 and 2.3 do not differ substantially between the two methods. However, when the true BFs grow larger, those computed with BayesRS tend to overestimate the evidence in favor of the Alternative. The overestimation seems to be slightly more pronounced for the smaller sample size. As in Study 2, we argue that the distortions are not problematic because in neither case they change the interpretation of the results of a given data set. Hence, the BayesRS package is able to discriminate between the Null and the Alternative of a fixed effect with associated random slopes. Although there is a substantial overestimation of the BF in some cases, this applies only when the BF in favor of the Alternative computed with the *t*-test is already large.

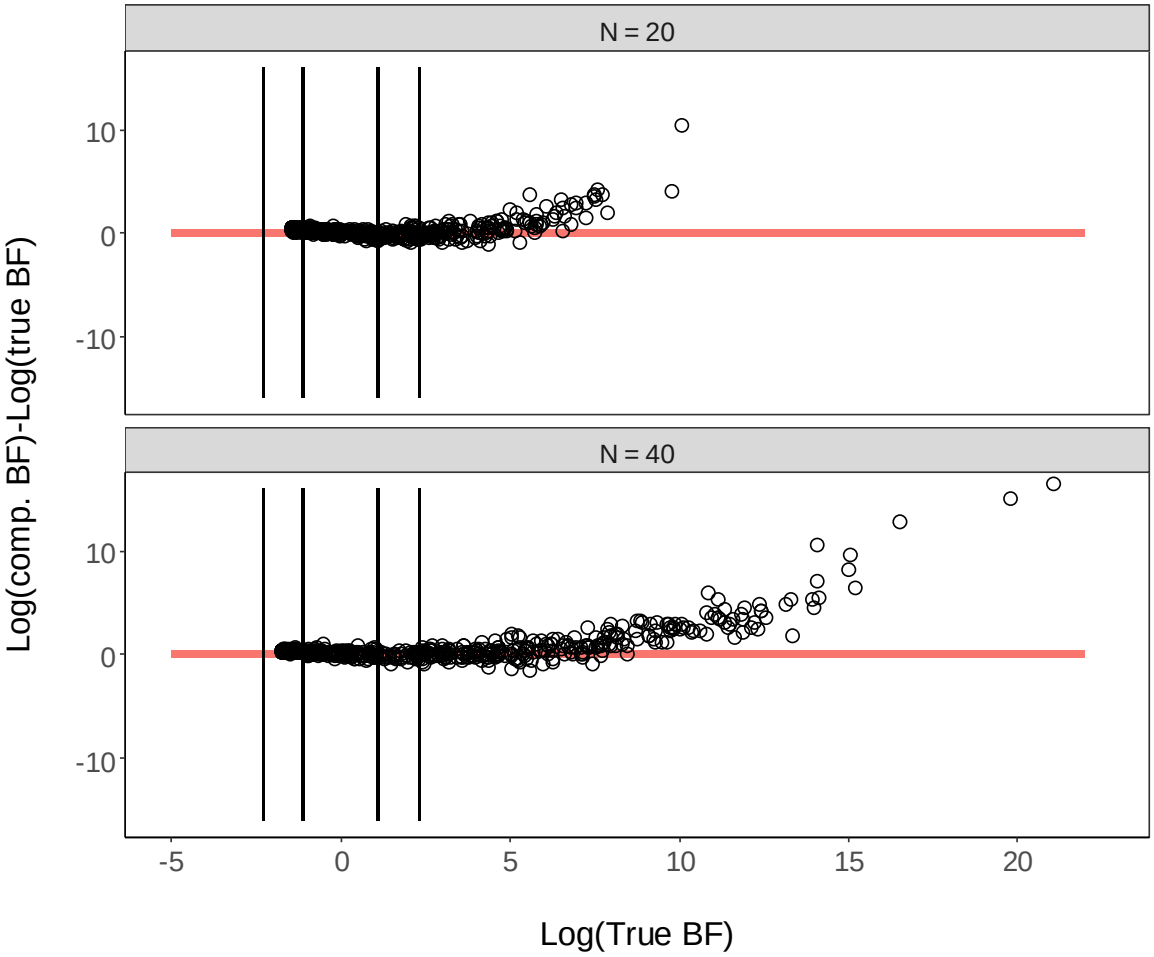

Figure 5. Deviation of the BF computed with BayesRS from the "true" BF from the Bayesian *t*-test computed with the BayesFactor package, plotted in Log space. The bold vertical lines correspond to BFs of 3.2 in favor of the Null (negative values) or the Alternative (positive values). The thin vertical lines correspond to BFs of 10 in favor of the Null (negative values) or the Alternative (positive values). A value on the red line indicates that the BF computed with BayesRS is the same as the true BF.

Simulation Study 4. Work by Barr and colleagues (2013) has shown that not accounting for true random slopes heavily increases Type 1 errors when *p*-values are used for inference. This may be problematic as there is no currently available tool that computes BFs and adequately accounts for the random-effects structure of continuous predictors. That is, using a Bayesian

model with only a fixed effect on the continuous predictor might overestimate the evidence in favor of the Alternative.

The concept of Type 1 and Type 2 errors does not fit well into the Bayesian framework. Bayesian inference – whether it is based on posterior distributions or based on BFs – does not involve accepting or rejecting a hypothesis, but rather assigns hypotheses a posterior degree of credibility. However, if the posterior degree of credibility of a hypothesis is overestimated, it leads our belief about the hypothesis in a wrong direction. This problem may even be aggravated when a decision criterion to decide between competing hypotheses is applied (e.g., does the 95 % HDI of a posterior exclude a region of practical equivalence, Kruschke & Liddell, 2017, or, do we stop data collection when the BF in favor of a hypothesis exceeds a threshold, Rouder, 2014). Therefore, we tested whether not accounting for actually present random slopes overestimates the evidence in favor of the Alternative when the Null is actually true (i.e., effect size on the fixed effect is zero). In the following, we refer to that overestimation as a "potential Type 1 error".

To this end, we again generated data from an experiment as in Study 1 with one continuous predictor and by-subject random slopes. We varied the number of subjects between 20, 40, and 60, the number of observations per design cell between 3, 6, and 9, and the effect size of the continuous predictor between 0, .15, .3, .45, .6, and .75. We fit the data once with only a by-subject random-intercepts model using the BayesFactor package that does not allow the specification of random slopes on continuous predictors, and once with a by-subject random-intercepts and random-slopes model with BayesRS. Specifically, we were interested to investigate what the models infer when the mean effect of the continuous variable is zero, as this will be informative for potential Type 1 errors. For illustrative purposes, we classified the computed BFs into the categories suggested by Kass and Raftery (1995): Strong Null ($BF < 0.1$), Substantial Null ($0.1 < BF < 0.3125$), Ambiguous ($0.3125 < BF < 3.2$), Substantial Alternative

(3.2 < BF < 10), and Strong Alternative (BF > 10). The results of Study 4 are shown in Figure 6 and Figure 7.

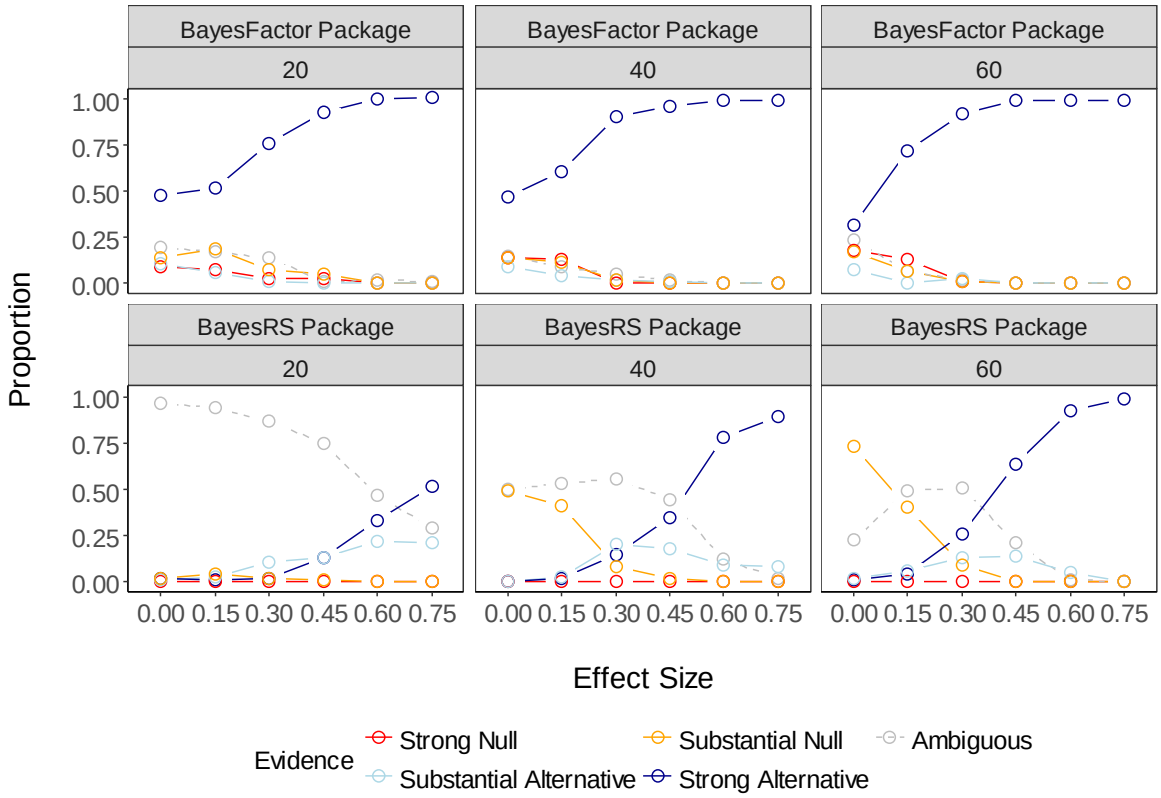

Figure 6. Proportion of BFs falling in one of the five categories suggested by Kass and Raftery (1995) plotted against the true effect size within the different design cells.

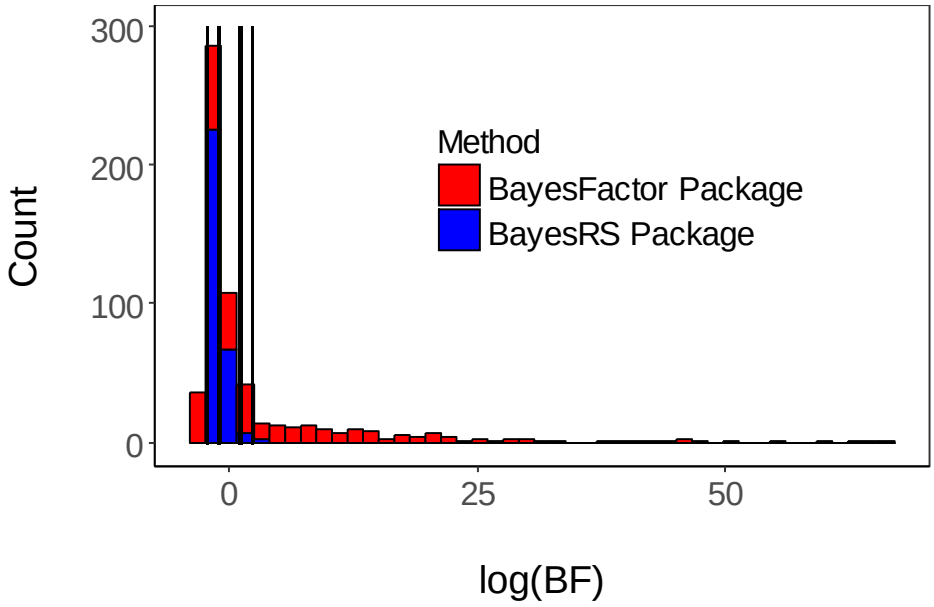

Figure 7. Distribution of log (BFs) when the effect size is 0. Note. The range of the x-axis differs between the two panels. The bold vertical lines correspond to BFs of 3.2 in favor of the Null (negative values) or the Alternative (positive values). The thin vertical lines correspond to BFs of 10 in favor of the Null (negative values) or the Alternative (positive values).

First, let us focus on the model without a random slope that was run with the BayesFactor package and is shown in the upper row of Figure 6. When effect size is zero, the model massively overestimates the evidence in favor of the alternative hypothesis. Although in this scenario the BF should favor the null hypothesis, in 42 % of the data sets the BF is regarded as strong evidence, and in a further 9 % as substantial evidence, in favor of the Alternative. In 20 % the BF is ambiguous, and in only 14 % and 15 % it shows strong and substantial evidence in favor of the Null, respectively. The red histogram in Figure 7 shows the distribution of BFs from the BayesFactor package when the true effect is zero. The long tail to the right reflects strong but wrong evidence in favor of the Alternative.

Now, let us focus on the bottom row in Figure 6 that shows the results of BayesRS. When focusing on effects of size zero, in only 1 % of the data sets the BF reflects strong evidence, and in a further 1 % substantial evidence in favor of the Alternative hypothesis. In 41 % of the data sets the BF reflects substantial evidence in favor of the Null, and in the remaining 57 % the evidence is ambiguous. The blue histogram in Figure 7 confirms that the log(BF) values from BayesRS are predominantly in the normative region < 0, and rarely stray much into the positive region. Hence, accounting for the random slopes decreases potential Type 1 errors when using BFs for inference, analogous to the conclusion of Barr et al. (2013) for frequentist mixed-effects models.

As can be seen in Figure 6, fewer potential Type 1 errors comes at the expense of less sensitivity to detect an effect in the model with a random slope. It takes more subjects and larger effect sizes for the BF to grow larger than 10 for the alternative hypothesis when it is true. This is, however, not an undesirable property because it essentially means that more data are needed to decide about the existence of small than of large effects. Although there is a chance of about 28 % for the BF to show substantial evidence in favor of the Null when the actual effect is .15, the BF is ambiguous in most of the cases, and it never shows strong evidence in favor of the Null. We argue that the increased potential Type 1 error rate in the model without random slopes is more problematic than the conservative nature of the model with random slopes. That is, in 42 % of the cases when there are true random slopes on the continuous predictor but the fixed effect is zero, researchers omitting random slopes from the model would conclude that there is strong evidence in favor of the effect. This argument weighs even more when it is considered that committing Type 1 errors has been regarded to be worse than committing Type 2 errors (Neyman, 1950). The results of this study therefore buttress the claim that random slopes have to

be accounted for, also for continuous predictors, in Bayesian as much as in frequentist mixed-effects models.

Simulation Study 5. In a last simulation study, we further explored a point put forward by Barr et al. (2013). Barr and colleagues observed that models that do not account for actually present correlations between random effects lead to slightly increased Type 1 errors in some circumstances, but also to slightly increased power. We incorporated a feature in BayesRS that allows users to model correlations between random effects. Here, we tested whether not accounting for actually present correlations between random slopes leads to increased potential Type 1 error rates, and to increased potential power when the BF is used for inference.

We simulated data from an experiment that slightly differed from the previous ones. In addition to a continuous predictor we now also included a categorical predictor. The by-subject slopes of the continuous variable and the categorical variable were drawn from a multivariate normal distribution. Thereby, (a) the mean effect of the continuous variable was fixed at zero to explore overestimation of the evidence (potential Type 1 error), but the mean effect of the categorical variable was either 0.2 or 1, also to explore a potential boost of the evidence for a true effect (i.e., potential power), (b) the correlation between the by-subject slopes of the two predictors was either 0.2, 0.5, or 0.8, and (c) the standard deviations of the by-subject slopes of both predictors were fixed to 1. This time, we simulated 100 data sets with 8 observations for every design cell of the simulated experiment and saved 100’000 MCMC steps for two experiments with 20 and 100 subjects, respectively.

First, let us focus on the continuous predictor for which there was no true mean effect, which is diagnostic for potential Type 1 errors. The distribution of BFs is plotted in Figure 8 for the two different model types (i.e., models with/without modeled correlations). As the results did not vary systematically, neither with the size of the correlation between the random slopes nor

with the size of the categorical effect, we pooled across these variables. Neither model leads to more than a negligible number of potential Type 1 errors. Across all generated data sets, the model without correlations led to 0.1 % of BFs larger than 10 (2 cases). The model with correlations led to 0.06 % of BFs larger than 10 across all generated data sets (1 case). Hence, accounting for the correlations only has a negligible effect on potential Type 1 error rate. It is clear that inference regarding the fixed effect of the continuous variable does not differ between the two models when the sample size is large (lower row). However, when the sample size is small (upper row) we can observe that the model without correlations is actually slightly more sensitive to detect a Null effect. In other words, the proportion of BFs that provide substantial evidence in favor of the Null is larger in the model without correlations than in the model with correlations, and this is at the expense of fewer BFs categorized as ambiguous in the model without correlations, corroborating the findings by Barr et al. (2013). Considering these results, we conclude that not accounting for actually present correlations between random slopes is not problematic, and may even be desirable when only data from a few subjects are available.

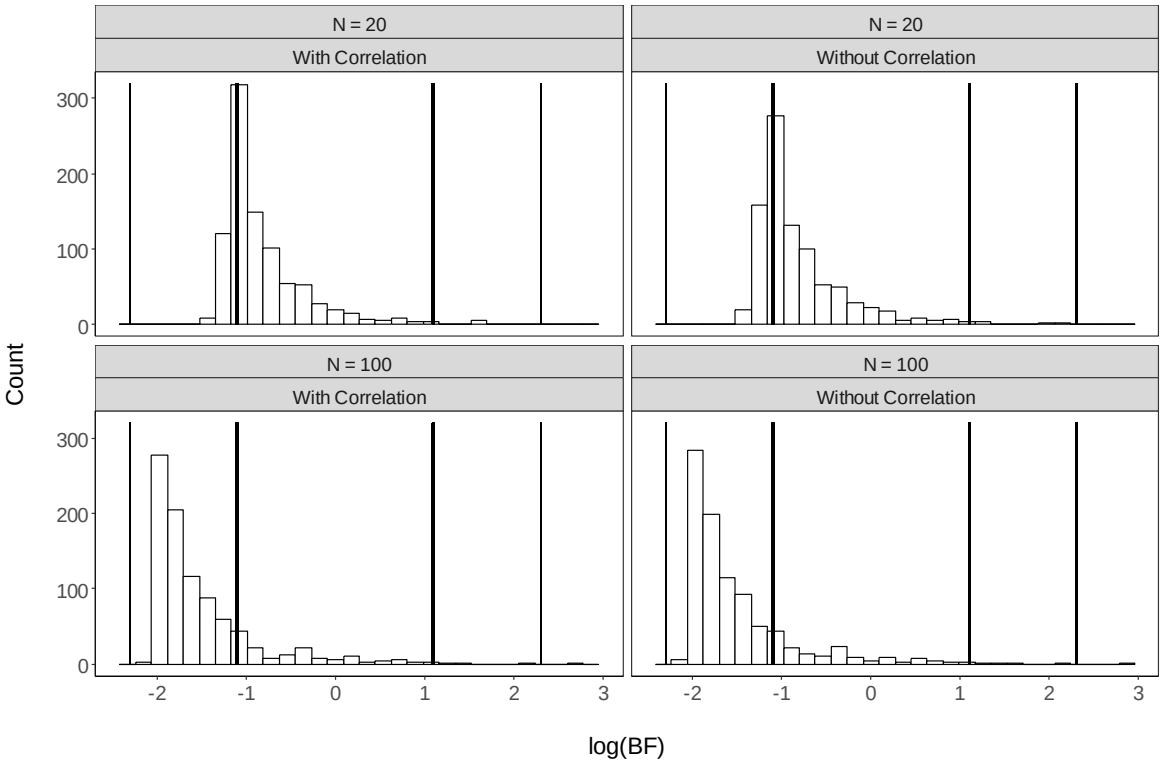

Figure 8. Distribution of log (BFs) for continuous predictor with true effect of 0 when accounting for the correlation (upper row) and when not accounting for the correlation (lower row). Note. The bold vertical lines correspond to BFs of 3.2 in favor of the Null (negative values) or the Alternative (positive values). The thin vertical lines correspond to BFs of 10 in favor of the Null (negative values) or the Alternative (positive values).

Second, let us focus on the categorical predictor whose effect size was varied between 0, .2, and 1, which is diagnostic for potential Type 2 errors. Density plots of the BFs are depicted in Figure 9. We again pooled across correlation size. Here, the differences between the two model types are slightly larger. The model without correlations leads to slightly decreased potential Type 2 error rates when the effect size of the categorical variable is small (i.e., .2) and when there are only 20 subjects in the experiment. This can be seen in the middle plot of the upper row: The proportion of BFs that are classified as ambiguous/substantial Null is smaller in the model without correlations than in the model with correlations, and instead the model without

correlations yielded more BFs reflecting substantial or strong evidence in favor of the alternative hypothesis. When there are 100 subjects in the experiment, this tendency is reduced but still present. Considering both potential Type 1 and potential Type 2 error rates, it appears unproblematic to ignore actually present correlations between random slopes. In contrast, the present results even suggest that a model that does not account for the correlation may be the better choice for determining whether or not a fixed effect is present. That is, while potential Type 1 error rates are largely the same, the model omitting correlations has a higher sensitivity to detect small effects in small samples, confirming for the Bayesian framework what Barr et al. observed in the frequentist framework.

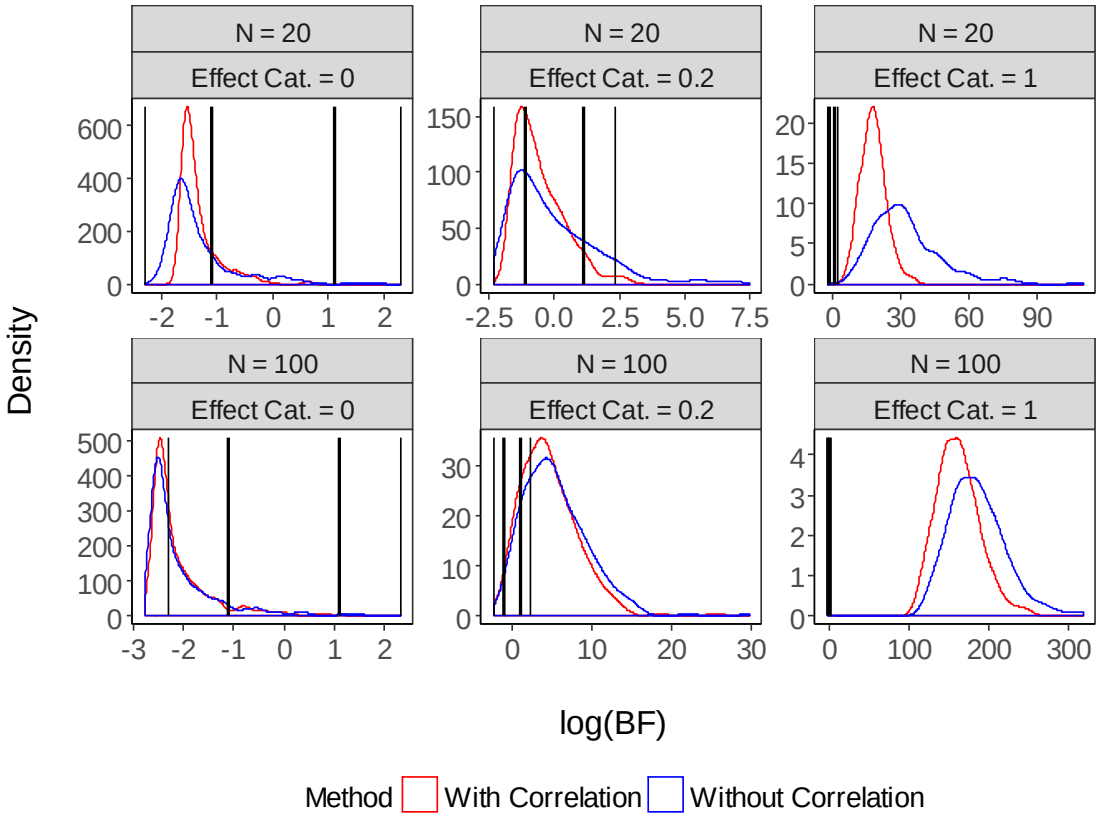

Figure 9. Density plots of log (BFs) within the different design cells. Note that the ranges of the y- and x-axes between panels differ. The bold vertical lines correspond to BFs of 3.2 in favor of the Null (negative values) or the Alternative (positive values). The thin vertical lines correspond to BFs of 10 in favor of the Null (negative values) or the Alternative (positive values).

7.5. General Discussion

Statisticians have advocated the use of Bayesian statistics for inference instead of *p* values, and the use of mixed-effects models or, more generally, hierarchical models. Mixed-effects models tend to overestimate the evidence in favor of the alternative hypothesis when they erroneously fail to include random slopes (Barr et al., 2013). As our Simulations 4 and 5 show, this is not only true for frequentist but also for Bayesian methods of inference on mixed-effects models. This raises a problem, because so far, there was no easy-to-use statistical tool that

computes BFs for models with random slopes on continuous predictors. Here, we closed this gap by introducing a new R package, BayesRS. The new package is also able to model correlations between random effects. We tested its functionality in five simulation studies.

The first study showed that the BF in favor of a fixed effect with associated random slopes can be estimated with good reliability. Especially BFs that are critical to decide whether the evidence in favor of an effect is not worth to be mentioned, substantial, or strong (Kass & Raftery, 1995) can be estimated with high reliability. When BFs grow larger their estimation becomes more variable. A BF in the region of several thousand may vary as much as about 20 %. However, usually it does not matter whether a BF is 3000 or 2400. Either value reflects decisive evidence in favor of a hypothesis. The second study showed that the BF of a fixed effect of a continuous predictor without associated random slopes is a good estimate of the true BF when it is small, but it slightly overestimates BFs when they grow large. In the third study, the BF of a fixed effect with associated random slopes was a good approximation of the BF of a Bayesian *t*-test on the true *b*-values. Again, BayesRS overestimated the BF when the BF of the *t*-test grew large. To summarize, the first three studies provide evidence that BayesRS estimates BFs that are reliable and valid (i.e., a good approximation of the true BF).

In two further studies (Simulations 4 and 5), we showed that not accounting for true random slopes overestimates the evidence in favor of the alternative hypothesis, and often signals strong evidence for it when it is actually not true, which echoes previous findings using *p* values for inference (Barr et al., 2013). Although there are clear differences between frequentist and Bayesian statistics, we argue that accounting adequately for the structure in the data is necessary regardless of the statistical framework. This means that if there are true differences in an effect between observation units but the differences are not accounted for in the structure of the

statistical model, a simplification of the structure will lead to biased inference. Simulation 5 also showed that modeling correlations between random effects leads to a slightly increased potential Type 2 error rate when both effect size and sample size are small. Therefore, when correlations between random effects are not of primary interest in a study, not accounting for them in the model may actually slightly increase the sensitivity to detect a true effect.

Whereas we strongly advocate including random slopes in a model, we do not advise to do so thoughtlessly, but rather we recommend to carefully consider how the data are structured. For example, the debate about how to set up an appropriate random-effects structure (Barr et al., 2013; Matuschek et al., 2017)highlights that researchers should carefully consider how their data are structured and inform their statistical models based on this consideration. Matuschek and colleagues suggest to test in a first step whether there is evidence for the random slopes in the data by using a fit criterion. If there is not, researchers could proceed omitting the random slopes. Further, the assumption that effects are distributed according to a normal distribution across units of observation is not going to be appropriate in all circumstances. For example, some subjects may be sensitive to an experimental manipulation, whereas others are not, which essentially leads to a bimodal distribution. To test such a hypothesis we incorporated an option in BayesRS that allows users to output the individual β values of an effect for all units of observation, so they can investigate their distribution. If the assumption of a normal distribution is not justified, researchers are encouraged to use a different model to analyze the data.

8. References

Alloway, T. P., & Alloway, R. G. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*(1), 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Awh, E., Barton, B., & Vogel, E. K. (2007). Visual Working Memory Represents a Fixed Number of Items Regardless of Complexity. *Psychological Science*, *18*(7), 622–628. https://doi.org/10.1111/j.1467-9280.2007.01949.x

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Special Issue: Emerging Data Analysis*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baddeley, A. (1996). The fractionation of working memory. *Proceedings of the National Academy of Sciences*, *93*(24), 13468–13472.

Baddeley, A. (2012). Working Memory: Theories, Models, and Controversies. *Annual Review of Psychology*, *63*(1), 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon Press.

Baddeley, A. D. (2001). Is working memory still working? *American Psychologist*, *56*(11), 851–864. https://doi.org/10.1037/0003-066X.56.11.851

Baddeley, A. D., & Hitch, G. J. (1974). The psychology of learning and motivation. *The Psychology of Learning and Motivation*, *8*, 47–90.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 575–589. https://doi.org/10.1016/S0022-5371(75)80045-4

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*(1), 158–173. https://doi.org/10.1037/0033-295X.105.1.158

Bailey, H., Dunlosky, J., & Kane, M. J. (2008). Why does working memory span predict complex cognition? Testing the strategy affordance hypothesis. *Memory & Cognition, 36*(8), 1383–1390. https://doi.org/10.3758/MC.36.8.1383

Bailey, H., Dunlosky, J., & Kane, M. J. (2011). Contribution of strategy use to performance on complex and simple span tasks. *Memory & Cognition, 39*(3), 447–461. https://doi.org/10.3758/s13421-010-0034-3

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time Constraints and Resource Sharing in Adults' Working Memory Spans. *Journal of Experimental Psychology: General, 133*(1), 83–100. https://doi.org/10.1037/0096-3445.133.1.83

Barrouillet, P., Bernardin, S., Portrat, S., Vergauwe, E., & Camos, V. (2007). Time and cognitive load in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 570–585. https://doi.org/10.1037/0278-7393.33.3.570

Barrouillet, P., Plancher, G., Guida, A., & Camos, V. (2013). Forgetting at short term: When do event-based interference and temporal factors have an effect? *Acta Psychologica, 142*(2), 155–167. https://doi.org/10.1016/j.actpsy.2012.12.003

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv Preprint arXiv:1406.5823*.

Bayes, T., Price, R., & Canton, J. (1763). *An essay towards solving a problem in the doctrine of chances*. C. Davis, Printer to the Royal Society of London.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Bower, G. H. (1970). Imagery as a relational organizer in associative learning. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 529–533. https://doi.org/10.1016/S0022-5371(70)80096-2

Bower, G. H., & Springston, F. (1970). Pauses as recoding points in letter series. *Journal of Experimental Psychology*, *83*(3, Pt.1), 421–430. https://doi.org/10.1037/h0028863

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Brener, R. (1940). An experimental investigation of memory span. *Journal of Experimental Psychology*, *26*(5), 467. https://doi.org/10.1037/h0061096

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576. https://doi.org/10.1037/0033-295X.114.3.539

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-Term Memory, Working Memory, and Executive Functioning in Preschoolers: Longitudinal Predictors of Mathematical Achievement at Age 7 Years. *Developmental Neuropsychology*, *33*(3), 205–228. https://doi.org/10.1080/87565640801982312

Camos, V. (2015). Storing Verbal Information in Working Memory. *Current Directions in Psychological Science*, *24*(6), 440–445. https://doi.org/10.1177/0963721415606630

Camos, V., & Barrouillet, P. (2014). Attentional and non-attentional systems in the maintenance of verbal information in working memory: the executive and phonological loops. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00900

Camos, V., Lagner, P., & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language*, *61*(3), 457–469. https://doi.org/10.1016/j.jml.2009.06.002

Camos, V., & Portrat, S. (2015). The impact of cognitive load on delayed recall. *Psychonomic Bulletin & Review*, *22*(4), 1029–1034. https://doi.org/10.3758/s13423-014-0772-5

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., … Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, *20*.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Chen, Z., & Cowan, N. (2005). Chunk Limits and Length Limits in Immediate Recall: A Reconciliation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1235–1249. https://doi.org/10.1037/0278-7393.31.6.1235

Chen, Z., & Cowan, N. (2009). Core verbal working-memory capacity: The limit in words retained without covert articulation. *The Quarterly Journal of Experimental Psychology*, *62*(7), 1420–1429. https://doi.org/10.1080/17470210802453977

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359. https://doi.org/10.1016/S0022-5371(73)80014-3

Conrad, R. (1964). Acoustic Confusions in Immediate Memory. *British Journal of Psychology*, *55*(1), 75–84. https://doi.org/10.1111/j.2044-8295.1964.tb00899.x

Cowan, N. (1999). An Embedded-Processes Model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). New York, NY, US: Cambridge University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(1), 87–114; discussion 114–185.

Cowan, N., Saults, J. S., Elliott, E. M., & Moreno, M. V. (2002). Deconfounding Serial Recall. *Journal of Memory and Language*, *46*(1), 153–177. https://doi.org/10.1006/jmla.2001.2805

Cowan, N., Zhijian Chen, & Rouder, J. N. (2004). Constant Capacity in an Immediate Serial-Recall Task. *Psychological Science (Wiley-Blackwell)*, *15*(9), 634–640. https://doi.org/10.1111/j.0956-7976.2004.00732.x

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Craik, F., & Tulving, E. (1975). Deth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294.

Daneman, M., & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(4), 561–584. https://doi.org/10.1037/0278-7393.9.4.561

Dickey, J. M., & Lientz, B. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*(1), 214–226.

Dunlosky, J., & Kane, M. J. (2007). The contributions of strategy use to working memory span: A comparison of strategy assessment methods. *The Quarterly Journal of Experimental Psychology*, *60*(9), 1227–1245. https://doi.org/10.1080/17470210600926075

Ecker, U. K. H., Oberauer, K., & Lewandowsky, S. (2014). Working memory updating involves item-specific removal. *Journal of Memory and Language*, *74*, 1–15. https://doi.org/10.1016/j.jml.2014.03.006

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://doi.org/10.1037/0033-295X.102.2.211

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(2), 223–271. https://doi.org/10.1037/a0027371

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review*, *9*(1), 59–79. https://doi.org/10.3758/BF03196257

Garavan, H. (1998). Serial attention within working memory. *Memory & Cognition*, *26*(2), 263–276. https://doi.org/10.3758/BF03201138

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis* (Vol. 2). Chapman & Hall/CRC Boca Raton, FL, USA.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. Retrieved from https://www.mpib-berlin.mpg.de/volltexte/institut/dok/full/gg/ggstehfda/ggstehfda.html

Gilchrist, A. L., & Cowan, N. (2011). Can the focus of attention accommodate multiple, separate items? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1484–1502. https://doi.org/10.1037/a0024352

Gobet, F., & Clarkson, G. (2004). Chunks in expert memory: Evidence for the magical number four … or is it two? *Memory*, *12*(6), 732–747. https://doi.org/10.1080/09658210344000530

Greene, R. L. (1987). Effects of maintenance rehearsal on human memory. *Psychological Bulletin*, *102*(3), 403–413. https://doi.org/10.1037/0033-2909.102.3.403

Grillon, M.-L., Johnson, M. K., Krebs, M.-O., & Huron, C. (2008). Comparing effects of perceptual and reflective repetition on subjective experience during later recognition memory. *Consciousness and Cognition*, *17*(3), 753–764. https://doi.org/10.1016/j.concog.2007.09.004

Guttentag, R. E. (1984). The mental effort requirement of cumulative rehearsal: A developmental study. *Journal of Experimental Child Psychology*, *37*(1), 92–106. https://doi.org/10.1016/0022-0965(84)90060-2

Hubbard, R. (2004). Alphabet Soup: Blurring the Distinctions Betweenp's anda's in Psychological Research. *Theory & Psychology*, *14*(3), 295–327. https://doi.org/10.1177/0959354304043638

Hulme, C., Maughan, S., & Brown, G. D. A. (1991). Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of Memory and Language*, *30*(6), 685–701. https://doi.org/10.1016/0749-596X(91)90032-F

Hulme, C., Roodenrys, S., Brown, G. D. A., & Mercer, R. (1995). The role of long-term memory mechanisms in memory span. *British Journal of Psychology*, *86*(4), 527.

Johnson, M. K., Reeder, J. A., Raye, C. L., & Mitchell, K. J. (2002). Second Thoughts Versus Second Looks: An Age-Related Deficit in Reflectively Refreshing Just-Activated Information. *Psychological Science (Sage Publications Inc.)*, *13*(1), 64.

Johnson, M. R., Higgins, J. A., Norman, K. A., Sederberg, P. B., Smith, T. A., & Johnson, M. K. (2013). Foraging for Thought An Inhibition-of-Return-Like Effect Resulting From Directing Attention Within Working Memory. *Psychological Science*, *24*(7), 1104–1112. https://doi.org/10.1177/0956797612466414

Johnston, J. C., McCann, R. S., & Remington, R. W. (1995). Chronometric evidence for two types of attention. *Psychological Science, 6*(6), 365–369.

Jones, T., & Oberauer, K. (2013). Serial-position effects for items and relations in short-term memory. *Memory, 21*(3), 347–365. https://doi.org/10.1080/09658211.2012.726629

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773–795. https://doi.org/10.2307/2291091

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 1–29. https://doi.org/10.3758/s13423-016-1221-4

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14*(4), 389–433. https://doi.org/10.1016/S0160-2896(05)80012-1

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge ; New York: Cambridge University Press.

Lewandowsky, S. (1999). Redintegration and Response Suppression in Serial Recall: A Dynamic Network Model. *International Journal of Psychology, 34*(5/6), 434–446. https://doi.org/10.1080/002075999399792

Lewandowsky, S., & Oberauer, K. (2015). Rehearsal in serial recall: An unworkable solution to the nonexistent problem of decay. *Psychological Review, 122*(4), 674–699. https://doi.org/10.1037/a0039684

Loaiza, V. M., Duperreault, K. A., Rhodes, M. G., & McCabe, D. P. (2014). Long-term semantic representations moderate the effect of attentional refreshing on episodic memory. *Psychonomic Bulletin & Review*, 1–7. https://doi.org/10.3758/s13423-014-0673-7

Loaiza, V. M., & McCabe, D. P. (2012). Temporal–contextual processing in working memory: Evidence from delayed cued recall and delayed free recall tests. *Memory & Cognition*, *40*(2), 191–203. https://doi.org/10.3758/s13421-011-0148-2

Loaiza, V. M., McCabe, D. P., Youngblood, J. L., Rose, N. S., & Myerson, J. (2011). The influence of levels of processing on recall from working memory and delayed recall tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1258–1263. https://doi.org/10.1037/a0023923

Logie, R. H., Cocchini, G., Sala, S. Della, & Baddeley, A. D. (2004). Is There a Specific Executive Capacity for Dual Task Coordination? Evidence From Alzheimer's Disease. *Neuropsychology*, *18*(3), 504–513. https://doi.org/10.1037/0894-4105.18.3.504

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, *10*(4), 325–337.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

Mayzner, M. S., & Schoenberg, K. M. (1964). Single-letter and digram frequency effects in immediate serial recall. *Journal of Verbal Learning and Verbal Behavior*, *3*(5), 397–400. https://doi.org/10.1016/S0022-5371(64)80008-6

Mora, G., & Camos, V. (2013). Two Systems of Maintenance in Verbal Working Memory: Evidence from the Word Length Effect. *PLOS ONE*, *8*(7), e70026. https://doi.org/10.1371/journal.pone.0070026

Mora, G., & Camos, V. (2015). Dissociating rehearsal and refreshing in the maintenance of verbal information in 8-year-old children. *Developmental Psychology, 6,* 11. https://doi.org/10.3389/fpsyg.2015.00011

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

Morey, R. D., & Rouder, J. N. (2014). BayesFactor: Computation of Bayes factors for common designs (Version 0.9.7). Retrieved from http://cran.at.r-project.org/web/packages/BayesFactor/index.html

Naveh-Benjamin, M., & Jonides, J. (1984). Maintenance rehearsal: A two-component analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(3), 369–385. https://doi.org/10.1037/0278-7393.10.3.369

Navon, D. (1984). Resources—a theoretical soup stone? *Psychological Review, 91*(2), 216–234. https://doi.org/10.1037/0033-295X.91.2.216

Navon, D., & Miller, J. (2002). Queuing or Sharing? A Critical Evaluation of the Single-Bottleneck Notion. *Cognitive Psychology, 44*(3), 193–251. https://doi.org/10.1006/cogp.2001.0767

Neyman, J. (1950). First course in probability and statistics. *First Course in Probability and Statistics, by J. Neyman. Published by Henry Holt, 1950.*

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*(1), 97–131.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 411–421. https://doi.org/10.1037/0278-7393.28.3.411

Oberauer, K. (2003). Understanding serial position curves in short-term recognition and recall. *Journal of Memory and Language*, *49*(4), 469–483. https://doi.org/10.1016/S0749-596X(03)00080-9

Oberauer, K. (2009). Chapter 2 Design for a Working Memory. In Brian H. Ross (Ed.), *Psychology of Learning and Motivation* (Vol. Volume 51, pp. 45–100). San Diego, CA: Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/S007974210951002X

Oberauer, K., & Bialkova, S. (2009). Accessing information in working memory: Can the focus of attention grasp two elements at the same time? *Journal of Experimental Psychology: General*, *138*(1), 64.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What Limits Working Memory Capacity? *Psychological Bulletin*. https://doi.org/10.1037/bul0000046

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review*, *115*(3), 544–576. https://doi.org/10.1037/0033-295X.115.3.544

Oberauer, K., & Lewandowsky, S. (2013). Evidence against decay in verbal working memory. *Journal of Experimental Psychology: General*, *142*(2), 380–411. https://doi.org/10.1037/a0029588

Oberauer, K., & Lewandowsky, S. (2014). Further evidence against decay in working memory. *Journal of Memory and Language*, *73*, 15–30. https://doi.org/10.1016/j.jml.2014.02.003

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review*, *19*(5), 779–819. https://doi.org/10.3758/s13423-012-0272-4

Page, M. P. A., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *364*(1536), 3737–3753. https://doi.org/10.1098/rstb.2009.0173

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220–244. https://doi.org/10.1037/0033-2909.116.2.220

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366

Phaf, R. H., & Wolters, G. (1993). Attentional Shifts in Maintenance Rehearsal. *The American Journal of Psychology*, *106*(3), 353–382. https://doi.org/10.2307/1423182

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2014). R Core Team (2014) nlme: linear and nonlinear mixed effects models. R package version 3.1-117. *Available at H ttp://CRAN. R-Project. Org/package= Nlme*.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling (Vol. 124, p. 125). Presented at the Proceedings of the 3rd international workshop on distributed statistical computing, Vienna.

Plummer, M. (2013). rjags: Bayesian graphical models using MCMC. *R Package Version*, *3*.

Plummer, M., Stukalov, A., & Denwood, M. (2015). rjags: Bayesian Graphical Models using MCMC (Version 4-4). Retrieved from https://cran.r-project.org/web/packages/rjags/index.html

Portrat, S., Guida, A., Phénix, T., & Lemaire, B. (2016). Promoting the experimental dialogue between working memory and chunking: Behavioral data and simulation. *Memory & Cognition*, *44*(3), 420–434. https://doi.org/10.3758/s13421-015-0572-9

Portrat, S., & Lemaire, B. (2014). Is Attentional Refreshing in Working Memory Sequential? A Computational Modeling Approach. *Cognitive Computation*, 1–13. https://doi.org/10.1007/s12559-014-9294-8

Raye, C. L., Johnson, M. K., Mitchell, K. J., Greene, E. J., & Johnson, M. R. (2007). Refreshing: A Minimal Executive Function. *Cortex*, *43*(1), 135–145. https://doi.org/10.1016/S0010-9452(08)70451-9

R Development Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. R Foundation for Statistical Computing, Vienna, Austria.

Rerko, L., & Oberauer, K. (2013). Focused, unfocused, and defocused information in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1075–1096. https://doi.org/10.1037/a0031172

Richardson, J. T. E., & Baddeley, A. D. (1975). The effect of articulatory suppression in free recall. *Journal of Verbal Learning and Verbal Behavior*, *14*(6), 623–629. https://doi.org/10.1016/S0022-5371(75)80049-1

Rose, N. S., Myerson, J., Roediger, H. L. I., & Hale, S. (2010). Similarities and differences between working memory and long-term memory: Evidence from the levels-of-processing span task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 471–483. https://doi.org/10.1037/a0018405

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001

Rundus, D., & Atkinson, R. C. (1970). Rehearsal processes in free recall: A procedure for direct observation. *Journal of Verbal Learning and Verbal Behavior*, *9*(1), 99–105. https://doi.org/10.1016/S0022-5371(70)80015-9

Saito, S., Logie, R. H., Morita, A., & Law, A. (2008). Visual and phonological similarity effects in verbal immediate serial recall: A test with kanji materials. *Journal of Memory and Language*, *59*(1), 1–17. https://doi.org/10.1016/j.jml.2008.01.004

Schwibbe, M. H. (n.d.). Der Semantische Atlas. Retrieved from http://kulturkontor-goe.de/semat/semat.htm

Service, E. (1998). The Effect of Word Length on Immediate Serial Recall Depends on Phonological Complexity, Not Articulatory Duration. *The Quarterly Journal of Experimental Psychology Section A*, *51*(2), 283–304. https://doi.org/10.1080/713755759

Souza, A. S., & Oberauer, K. (2017). The contributions of visual and central attention to visual working memory. *Attention, Perception, & Psychophysics*, *79*(7), 1897–1916. https://doi.org/10.3758/s13414-017-1357-y

Souza, A. S., Rerko, L., & Oberauer, K. (2015). Refreshing memory traces: thinking of an item improves retrieval from visual working memory. *Annals of the New York Academy of Sciences*, *1339*(1), 20–31. https://doi.org/10.1111/nyas.12603

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583.

Süß, H.-M., Oberauer, K., Wittmann, W. W., Wilhelm, O., & Schulze, R. (2002). Working-memory capacity explains reasoning ability—and a little bit more. *Intelligence*, *30*(3), 261–288. https://doi.org/10.1016/S0160-2896(01)00100-3

Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review*, *15*(3), 535–542. https://doi.org/10.3758/PBR.15.3.535

Team, R. C. (2017). R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2014.

Thalmann, M., Niklaus, M., & Oberauer, K. (2017). Estimating Bayes Factors for Linear Models with Random Slopes on Continuous Predictors. *PsyArXiv*. https://doi.org/10.17605/OSF.IO/4XQVR

Thalmann, M., & Oberauer, K. (2017). Domain-specific interference between storage and processing in complex span is driven by cognitive and motor operations. *The Quarterly Journal of Experimental Psychology*, *70*(1), 109–126. https://doi.org/10.1080/17470218.2015.1125935

Tombu, M., & Jolicœur, P. (2003). A central capacity sharing model of dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(1), 3–18. https://doi.org/10.1037/0096-1523.29.1.3

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*(2), 127–154. https://doi.org/10.1016/0749-596X(89)90040-5

Vergauwe, E., Camos, V., & Barrouillet, P. (2014). The Impact of Storage on Processing: How Is Information Maintained in Working Memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/a0035779

Vergauwe, E., & Cowan, N. (2014). A common short-term memory retrieval rate may describe many cognitive procedures. *Frontiers in Human Neuroscience*, *8*, 126. https://doi.org/10.3389/fnhum.2014.00126

Vergauwe, E., & Langerock, N. (2017). Attentional refreshing of information in working memory: Increased immediate accessibility of just-refreshed representations. *Journal of*

Memory and Language, *96*(Supplement C), 23–35. https://doi.org/10.1016/j.jml.2017.05.001

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems ofp values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, *60*(3), 158–189.

Wagenmakers, E.-J., Van Der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*(1), 3–22. https://doi.org/10.3758/BF03194023

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, *16*(4), 752–760. https://doi.org/10.3758/PBR.16.4.752

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860

Curriculum Vitae

MIRKO THALMANN

mirkothalmann@hotmail.com

SWISS, BORN ON 1987-12-03

Education

03/2014 – 11/2017 **Student of PhD Program in Psychology**

- University of Zurich, Department of Psychology (Cognitive Psychology Unit)
  - Dissertation title: “Chunking and Rehearsal in Working Memory: A Matter of Central Attention?”
- Committee: Prof. Dr. Klaus Oberauer (University of Zurich), Prof. Dr. Edward Awh (University of Chicago)

09/2011 – 01/2014 **Master of Science in Psychology (Minor in Law)**

- University of Zurich
- Area of specialization: Cognitive Psychology and Cognitive Neuropsychology
- Cumulative degree: 5.7 (scale from 1-6, at which 6 is the best)
- Master thesis title: “Domain-Specific Interference in Working Memory under Control of Cognitive Load”

09/2007 – 07/2010 **Bachelor of Science in Psychology (Minors in Law and Political Science)**

- University of Zurich
- Cumulative degree: 5.3 (scale from 1-6, at which 6 is the best)

08/2001 – 07/2006 **Matriculation examination (Matura)**

- Cantonal School, St. Gallen

Teaching and Advising

Fall term 2015 **Seminar Lecturer "Debates in Cognitive Psychology"**

- As PhD student at the University of Zurich
- Lecturing sessions of the seminar
- Mentoring students in preparation and presentation of controversial debates
- Acting as a moderator in the discussions of the opposing teams

03/2014 up to the present (2 years) **Supervisor and Mentor of student assistants and interns**

- As PhD student at the University of Zurich
- Guiding an intern in programming experiments and analyzing data
- Supervision of several student assistants in project realization

Peer-Reviewed Articles as First Author

- Thalmann, M., & Oberauer, K. (2017). Domain-specific interference between storage and processing in complex span is driven by cognitive and motor operations. *The Quarterly Journal of Experimental Psychology*, 70(1), 109–126.
- Thalmann, M., Souza, A. S., Oberauer, K. (Manuscript accepted for publication in *Journal of Experimental Psychology: Learning, Memory, and Cognition*). How Does Chunking Help Working Memory?
- Thalmann, M., Souza, A. S., Oberauer, K. (Manuscript submitted for publication). Revisiting the Attentional Costs of Rehearsal in Working-Memory Tasks.
- Thalmann, M., Niklaus, M., & Oberauer, K. (2017, October 18). Estimating Bayes Factors for Linear Models with Random Slopes on Continuous Predictors. http://doi.org/10.17605/OSF.IO/7ZPK3 (Manuscript submitted for publication)

Grants

01/17-08/17

- SNSF Doc.Mobility Grant "Building up Long-Term Memory of visually presented Configurations – Resolving Discrepancies in the Literature". University of New South Wales (Sydney, Australia)

Acknowledgements

- Thanks to Klaus Oberauer for his close supervision of my PhD thesis and for creating a welcoming research atmosphere in which questions are considered to be stimulating and not annoying as elsewhere. Many thanks also for bringing me into contact with Chris Donkin. At the same time, I would like to thank Ed Awh for being the second supervisor of my thesis.

- Thanks to Alessandra Souza for valuable input in many discussions and for delivering proper liquor from Brazil.

- I would also like to thank all present and former members of the UZH cognitive psychology unit for good conversations about research and for the good atmosphere in the lab. Special shout outs to Carla, Lea, Marcel, and Mirjam for the great company, sometimes good lunches not in the Mensa, and interesting conversations.

- Thanks to Chris Donkin for his supervision during my research stay and cheers to the UNSW 721 crew for making my eight months in Sydney a great experience.

- I am deeply grateful to my parents. Without their constant support during my Bachelor, Master, and PhD studies, this thesis would not have been possible. Thank you!

- Finally, I would like to especially thank Cornelia for constant help to make my research understandable to a broader audience.